

Reaching a tipping point

A popular new paradigm for the nature of change pertains more to the social and political worlds than it does to the physical one.

Rarely since *Catch 22* has a book title made its mark on the language in such a way as *The Tipping Point*. The author, Malcolm Gladwell, made no claim to have invented the term, but his thought-provoking book brought the idea into common parlance. Since then, the view that incremental changes in a cause can suddenly produce a much larger effect has entered common currency. It is now being ever more frequently deployed in the debate about the world's climate.

In some respects, this is old wine in new bottles. For almost as long as people have been worrying about anthropogenic climate change, there have been warnings that, although the build-up of greenhouse gases may be slow and gradual, the effects they will have on the system need not be. The physical, chemical and biological responses that turn greenhouse gases into climate change are complex and subtle, and capable of responses that are surprisingly disproportionate. There are thresholds beyond which the past response of the system no longer predicts the future, and there are positive feedbacks through which change can feed on itself. All these possibilities are now being discussed under the rubric of tipping points.

It is reasonable to worry about such things, but there are three dangers attendant on focusing humanity's response to the climate crisis too much on tipping points. The first is the uncertainty of the science; the second is the tendency of such an emphasis to distort our responses; the third is the danger of fatalism.

The models through which our understanding of the climate system are channelled into assessments of how it might behave in the future are impressive by the standards of human investigation, but crude with respect to the details of the Earth system. All sorts of phenomena, from the formation of clouds to the respiration of soils, are hard to capture accurately, and it is on such details that an understanding of possible tipping points depends (see page 802). Anyone

claiming to know for sure when a particular tipping point will be reached should be treated with suspicion — and so must anyone who suggests that no tipping point will ever be reached.

The second problem is that an emphasis on tipping points not yet reached increases the focus on the future. Such an increase tips the balance away from adapting to climate change and in favour of trying to avoid it. A rational response to the challenge of the twenty-first century's climate is to do both: to reduce the rate at which greenhouse gases force climate change, but at the same time build up the ability to cope with adverse climates.

The third issue is that tipping points can induce fatalism. The concept may encourage the belief that a complete solution is the only worthwhile one, as any other course may allow the climate system to tumble past the crucial threshold. This sort of all-or-nothing approach is already over-stressed in climate policy by the Framework Convention

on Climate Change, which calls for the complete avoidance of dangerous anthropogenic climate change, rather than the more reasonable and more feasible goal of minimizing and controlling it.

The concept of the tipping point is, in fact, more pertinent to the climate crisis in the social sphere than in the physical world. The strength of Gladwell's book lies in its reasoned illustrations of the ways in which beliefs and behaviours change, and the rules and contexts that govern that change. It is possible to make people change their minds and behaviours, and for those changes to spread like a contagion. "Look at the world around you," Gladwell argues. "It may seem like an immovable, implacable place. It is not. With the slightest push — in just the right place — it can be tipped." ■

"Anyone claiming to know for sure when a particular tipping point will be reached should be treated with suspicion."

A fresh start

Will a change of management at Los Alamos put basic research under pressure?

Earlier this month, with mercifully little fanfare, scientists at Los Alamos National Laboratory in New Mexico witnessed a changing of the guard. For 63 years, the world's best-known nuclear-weapons laboratory had been under the management of the University of California. From 1 June, the university will run it in partnership with three commercial companies.

With luck, the change will conclude a period of turbulence at the laboratory, which began when Taiwanese-born scientist Wen Ho Lee was suspected of espionage in 1999 (see *Nature* 398, 96; 1999). Lee

was never charged, and in the end pleaded guilty to mishandling classified data, but the lab has since been plagued by a series of security and procurement scandals. That encouraged the Department of Energy, which oversees the facility, to open its management contract up to bidding for the first time (see *Nature* 423, 104; 2003).

Under a contract that runs until 2013, the university will run the laboratory jointly with San Francisco-based Bechtel National, an engineering contractor that already manages several US nuclear weapons facilities. Two other companies, BWX Technologies and Washington Group International, will also play roles in managing the laboratory's nuclear equipment and its hazardous waste.

The energy department and its congressional overseers hope the new arrangement will bring best business practice to the laboratory and draw its recent difficulties to a close. The laboratory's staff, who will retain their pension rights and status within the University of



California, are also hoping that the change will quell public criticism of the laboratory.

But for scientists at the lab, this hope is tinged with concern that the new management will shift its focus away from basic research. Both BWX Technologies and Bechtel have a track record in engineering management, overseeing production facilities for US nuclear weapons. They are for-profit institutions, more interested in meeting government schedules than nurturing scientific excellence.

There are still grounds for optimism that science won't take a back seat at Los Alamos. The lab's new director, Michael Anastasio, is a physicist with a distinguished career in nuclear-weapons design. Other Department of Energy facilities under joint business and university management, such as Oak Ridge National Laboratory in Tennessee, have continued to do good science.

None of these places enjoy Los Alamos's intellectual reputation, however, and fears persist of a shift in priorities at the New Mexico laboratory. The energy department is considering the construction of a multi-billion-dollar facility there to produce replacement plutonium triggers for existing nuclear weapons. In the zero-sum game

that is the Department of Energy's budget process, scientists fear that a larger weapons-production role at Los Alamos will come at the expense of its research programmes — especially those not directly related to nuclear weapons.

The main strength of Los Alamos has always been its ability to attract good people by offering exceptional facilities dedicated to a range of scientific enquiry. It is home to some of the world's most powerful supercomputers, which are used for such purposes as simulating a possible bird-flu epidemic. The laboratory also houses acknowledged global leaders in neutron science, mathematics, computer science, nuclear non-proliferation, and more. The scientists and engineers who work on nuclear weapons there rely on the non-weapons work to retain links with the wider academic world. That interface is what makes Los Alamos special, and the laboratory's fresh management team should nurture it with care.

"Scientists fear that a larger weapons-production role at Los Alamos will come at the expense of its research programmes."

Cash-per-publication...

.....is an idea best avoided.

South Korea has become the latest country to offer scientists cash prizes for publications in top-level international journals (see page 792). Other nations, including China and Pakistan, already have such programmes in place. The thousands, or even tens of thousands, of dollars on offer can be a fat prize for researchers in countries with lean science budgets.

Many researchers will turn up their noses at this practice. Scientists, after all, are supposed to be motivated by curiosity, by a devotion to finding the truth, by a desire to solve various philosophical or social problems — not by money. And funds should find their way to self-motivated scientists with a project deemed important. This assessment should be made by taking account of the project's feasibility, originality and scientific significance.

But this is no easy task in any country. In countries with inexperienced or understaffed scientific evaluation committees, it is almost impossible. Without the proper mechanism in place, funding can be diluted by equal distribution to all, or hijacked by projects approved on the basis of personal connections rather than sound science. Where there is little faith in review committees, giving money directly to those who have proven themselves might seem a beneficial, albeit imperfect, way of encouraging scientists.

Proponents can point to other potential advantages too. Scaling up bonuses for high-impact papers, as these programmes often do, might stem the urge to churn out quick papers in order to beef up a publication list. (Graduate students in China often need to have several publications for higher degrees.) And they also encourage scientists in countries with traditions of local-language publishing to think more internationally.

But there are some powerful arguments against the widespread adoption of the practice. Cash bonuses tied to specific publications

are likely to exacerbate corrupting tendencies in the scientific community. Debates over who should be included on author lists, and who should be the first author and the corresponding author, will surely get even more vicious when a chunk of money is on the line. A scientist struggling to meet a mortgage payment might be more willing to forgo a potentially fantastic result for a quick cash-earner. And a researcher measuring science in terms of dollars might even be more tempted to plagiarize or fabricate data.

In countries recently damaged by high-profile cases of scientific corruption, where it is all the more essential to develop a culture of integrity, the award of large sums of money for high-impact publications is even less desirable. The scientific world already places too much importance on high-impact journals in assessing individuals, and on the crude 'impact factor' in particular (see *Nature* **435**, 1003–1004; 2005). Papers published in journals with high impact factors do indeed tend to be significant papers, but a literal formula highly geared in their favour cannot do justice to the way science works.

"A researcher measuring science in terms of dollars might be more tempted to plagiarize or fabricate data."

Money does matter, of course. Scientifically developing countries need to compete for excellent scientists in an increasingly global marketplace. And scientists everywhere use their reputations, based on their latest and greatest papers, to negotiate raises, promotions or new jobs. So countries struggling to find ways of motivating their scientists need to reward outstanding researchers for good publications, perhaps by paying a bonus based on a peer review of achievements over at least a year.

But nations and agencies should avoid resorting to crude cash-per-paper incentives. They should instead attempt to be less formulaic and more considered in the ways they reward their scientists. And they should redouble their efforts to ensure that, in the midst of concerns about rewards, young scientists are committed above all to the ethical pursuit of scientific truths.

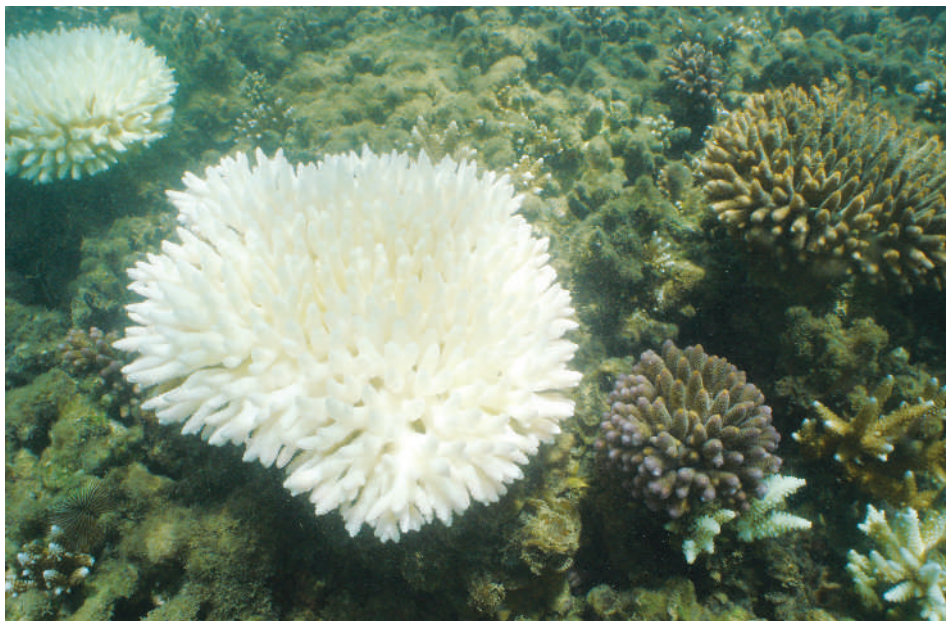
RESEARCH HIGHLIGHTS

In hot water*Proc. R. Soc. Lond. B*

doi:10.1098/rspb.2006.3567 (2006)

Warming waters threaten the intimate relationship between stony corals and the nutrient-providing zooxanthella algae that live within their tissues. But research from the Australian Institute of Marine Science, Townsville, offers some hope.

Ray Berkelmans and Madeleine van Oppen transplanted Great Barrier Reef coral colonies from cooler reefs to warmer inshore bays, where the corals expelled their symbiotic algae, a process known as coral bleaching (see white colonies, pictured). Some of the corals that recovered from the first bleaching event could then tolerate temperatures 1–1.5 °C warmer than previously. The researchers show that these corals had changed their algae to a more heat-resistant form.



R. BERKELMANS

GEOLOGY**After the flood***Geology* **34**, 461–464 (2006)

What caused the mass extinction at the end of the Early Cambrian epoch, about 508 million years ago?

Linda Glass and David Phillips of the Australian National University in Canberra have identified a new suspect: a massive outpouring of volcanic rock in what is now northern Australia. They have used radioisotope dating to fix the age of this formation, christened the Kalkarindji 'flood basalts' province, to around the time of the Cambrian mass extinction. Flood basalts elsewhere have been linked to other mass extinctions because of their effect on global climate and atmospheric chemistry — this is the oldest such association.

CELL BIOLOGY**Neurons and nematodes***Cell* **125**, 987–1001 (2006)

Scientists have identified a mechanism that plays a role in both a cell's response to stress and an organism's lifespan.

Azad Bonni of Harvard Medical School in Boston, Massachusetts, and his colleagues studied the response of mammalian neurons to oxidative stress. They found that FOXO transcription factors are the long-sought targets of a family of stress-activated enzymes, the MST kinases: the enzyme MST1 alters FOXO3, ultimately triggering the death of the damaged cell.

In nematode worms, the researchers implicated MST–FOXO signalling in ageing. Worms lacking the MST analogue CST-1 died young. Those with more of the enzyme had a longer lifespan, providing that the FOXO3 analogue was present.

MICROBIOLOGY**Hole punch***J. Cell Biol.* doi:10.1083/jcb.200509009 (2006)

Disease-causing *Staphylococcus aureus* bacteria (pictured below) may escape from the bloodstream by punching holes through the cells lining blood vessels, say researchers.

Emmanuel Lemichez at the National Institute of Health and Medical Research in Nice, France, and his colleagues studied the effects of an *S. aureus* toxin known as EDIN. EDIN blocks a protein called RhoA, which

acts rather like the foreman on a building site, telling cells where to construct their internal scaffolding in response to external signals. The loss of this coordination results in long tunnels, dubbed macroapertures, forming in the cell.

ORGANIC CHEMISTRY**Symmetry selective***J. Am. Chem. Soc.* doi:10.1021/ja061853f (2006)

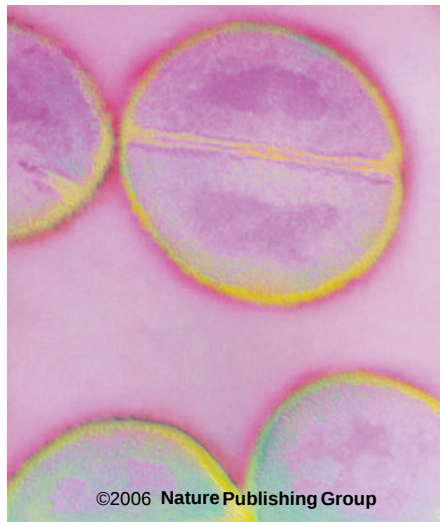
Dichroanone is part of a family of natural chemicals with anti-tumour properties. Now chemists have made one of its optical isomers from scratch — the first enantioselective synthesis of any chemical in this family.

Dichroanone includes three carbon rings fused to a central carbon atom in a configuration that is notoriously hard to build.

Ryan McFadden and Brian Stoltz of the California Institute of Technology, Pasadena, managed to produce (+)-dichroanone in 11 chemical steps with a 4% overall yield (pretty impressive for an asymmetric natural product synthesis). They avoid using protecting groups, which are normally needed to shield one part of a molecule while another segment is being transformed — a strategy that cuts the number of steps.

ASTRONOMY**Hydrogen hunt over***Astrophys. J.* **643**, 675–679 (2006)

Maps of nearby galaxies have resolved an astronomical riddle concerning the whereabouts of the material that fuels star



K. KIM/SPL

formation — molecular hydrogen.

The light from some quasars contains absorption lines, created when it passed through galaxies *en route* to Earth, which provide information about the galaxies' content. Atomic hydrogen shows up clearly in these systems, dubbed 'damped Lyman alpha absorbers', but there has been little evidence for molecular hydrogen.

Martin Zwaan of the European Southern Observatory and Jason Prochaska of the University of California's Lick Observatory, Santa Cruz, report that, in nearby galaxies, molecular hydrogen clusters densely in the galactic centre — occupying an area 150 times smaller than the atomic hydrogen. Hardly surprising, then, that light from quasars rarely passes through it.

MATERIAL SCIENCE

Polished finish

Science **312**, 1504–1508 (2006)

Round nanoparticles will give silicon a better polish, say researchers who have developed a recipe for cooking them up.

Abrasive pastes made from cerium oxide nanoparticles are used to flatten silicon wafers, a key step in the manufacture of integrated circuits. But such particles tend to have a faceted shape, with angular edges that can damage the wafer surface.

Xiangdong Feng, then at the Ferro Corporation in Independence, Ohio, and Zhong Lin Wang of the Georgia Institute of Technology in Atlanta synthesized their particles by setting a cerium spray alight. Adding titanium to this burning precursor mix gave nearly round particles, because the titanium forms a liquid shell around the crystallizing cerium oxide (pictured above).

ECOLOGY

The truth about trees

Science doi:10.1126/science.1124712 (2006)

A vast survey of the world's tropical forests is set to fuel the debate over what determines their diversity.

Richard Condit of the National Center for Ecological Analysis and Synthesis in Santa Barbara, California, and his colleagues analysed data from sites ranging from a dry forest in India, with just 73 species in 50 hectares, to a rainforest in Malaysia with 1,167 species in a similar-sized area.

It was previously thought that having tree species with a broad range of lifespans and growth rates would promote diversity, reasoning that a slow-growing tree, for example, would better compete with fast-growing neighbours if it were longer lived. Instead, the researchers found that the most diverse forests were the least varied demographically.

BIOLOGY

Three-way feast

PLoS Biol. doi:10.1371/journal.pbio.0040188 (2006)

The glassy-winged sharpshooter, *Homalodisca coagulata*, unlike many sap-sucking insects, saps from the water-carrying cells of the xylem, rather than the sugary sap of the phloem. Jonathan Eisen, then

at the Institute for Genomic Research in Rockville, Maryland, and his colleagues have now shown how a division of labour between the symbiotic bacteria living in the insect's

cells makes this low-nutrient lifestyle possible.

A complete genome of the bacterium *Baumannia cicadellinicola* and a partial sequence of *Sulcia muelleri* reveal that *Baumannia* has the metabolic pathways needed to

provide its insect host with vitamins, but has lost those for the essential amino acids. *Sulcia* devotes most of its very small genome to that neglected task, thus complementing its colleague. Each of the three organisms thus requires both of the others.

NANOTECHNOLOGY

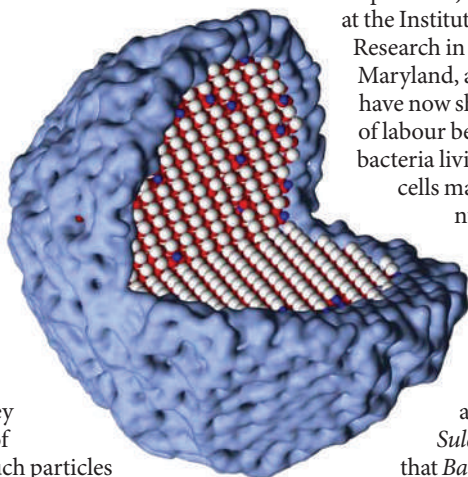
Trombone scales

Phys. Rev. Lett. **96**, 215503 (2006)

Tiny vibrating rods can be used as highly sensitive weighing scales, because an attached mass — down to a single virus particle — will change their oscillation frequency.

Researchers led by Alex Zettl at the University of California, Berkeley, have built such a device using a carbon nanotube that can be extended like a telescope. They say this allows them, "like a trombone player shifting notes", to tune the rod's vibrational frequency.

This should make the device sensitive to a wider range of masses than a tube of fixed length, as the resonant frequency can be matched to the mass to be measured.



D. C. SAYLE, X. D. FENG & Z. L. WANG

JOURNAL CLUB

Bart de Strooper
Flanders Interuniversity Institute
for Biotechnology and the
Catholic University of Leuven,
Belgium

An expert in Alzheimer's disease wonders why striking findings didn't hit the headlines.

Last year, *Nature Medicine* published a paper that, in my sincere naivety as a basic science researcher, I expected to prompt headlines such as 'Major crisis in pharmaceutical sector' and 'Stock owners dump shares'.

The paper (T. Kukar *et al.* *Nature Med.* **11**, 545–550; 2005) showed that a bunch of compounds, some registered as therapeutics, could augment in human cell cultures production of the small amyloid-beta peptide A β 42, believed to cause Alzheimer's disease.

The increases in A β 42 levels observed were quite dramatic, equaling or even surpassing those caused by the genetic mutations that lead to early and aggressive forms of genetic Alzheimer's.

To my surprise, the publication went largely unnoticed, even though many more modest papers get easily into the news. Recently, for instance, I was asked to comment on an ominous report claiming that intellectuals might be at higher risk of dementia. I felt immediately depressed, considering my secret long-term plans as emeritus professor. This insignificant story slipped easily out of my mind.

By contrast, the Kukar *et al.* paper leaves me uneasy. The authors, well aware of the potential implications of their observations, are careful in their extrapolations. However, they conclude: "exogenous compounds... may represent an unrecognized risk for the development of Alzheimer's disease".

I bet in many pharmaceutical companies the press officers were on red alert, while lawyers and managers spent precious time drafting a response to possible litigation because medication X could have caused Alzheimer's in grandad or grandma.

I wonder why this didn't happen.

NEWS

Science academies target G8 agenda

In a demonstration of how international negotiations should be done, 12 national science academies have issued two joint statements to the leaders of the G8 countries, who will meet at their annual summit in Russia next month. One endorses a reinvention of the world's disease surveillance system; the other calls for a major expansion of energy research to address a looming global crisis in energy supplies.

The statements were announced on 14 June by the academies of the G8 countries plus Brazil, China, India and South Africa. They follow the first such exercise at last year's Gleneagles summit in Britain, when Britain's Royal Society coordinated joint academy statements on climate change as well as capacity building for Africa.

The academies' stronger role in international advocacy is a "new and extremely positive development", says Martin Rees, president of the Royal Society. "It's a step towards the international scientific community having a more effective voice at the political level."

Rees believes the academies' input influenced the outcome of last year's G8, which

included greater debt relief for Africa. And he hopes the string of recommendations for disease surveillance and the energy crisis from the 12 academies (see box) will translate in firm pledges from this year's G8 meeting.

The academies argue that the size of global efforts in both infectious diseases and energy sourcing are out of touch with the scale of the problems. They lament the inadequacy of the current systems of national and international disease surveillance, which they describe as "multi-component and uncoordinated". The threat of avian flu, they argue, should be a catalyst for investment in a more tightly coordinated global system, that in particular would see animal and human health experts working more closely together.

Likewise, Rees says the G8 must address what he describes as serious inadequacies in funding and incentives for energy research: "In relation to the scale of the problem, the R&D effort worldwide is unduly low."

Although the G8 and other international political meetings are important for setting agendas and funding priorities, the academies

"It's a step towards the scientific community having a more effective voice at the political level."



recognize their shortcomings, says André Capron, foreign secretary of France's Académie des Sciences. In particular, he criticizes the "disappointing" habit of states in neglecting to honour pledges once they get home. At an avian-flu summit in Beijing in January, for example, countries pledged US\$1.9 billion in grants and loans to a global action plan, but so far only \$1 billion has been committed, and of that just \$286 million has been spent. Donors also often insist that funds

E. SUGITA/REUTERS

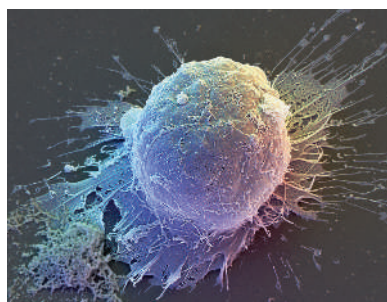
Koreans admit disguising stem-cell lines

Federally funded US researchers were supplied with a prohibited stem-cell line, disguised as an approved line by researchers at Seoul's MizMedi hospital. Allegations that a cell line had been switched surfaced earlier this month (see *Nature* 441, 680; 2006). Sung-il Roh, chairman of the board of trustees at MizMedi Women's Hospital, has now admitted to *Nature* that this was indeed the case.

Research with the unapproved cell line has been halted. But the incident highlights how hard it is for researchers to know what they are working with. And in this case the problem was not only scientific but

political: US president George W. Bush announced in 2001 that because of ethical concerns over human embryonic stem cells, only such cell lines established before 9 August 2001 would be eligible for federally funded research. One line established by researchers at MizMedi Hospital, Miz-hES1, was created before that date, and approved by the US National Institutes of Health (NIH). But the cell line sent out in its place, Miz-hES5, was created later.

The NIH has paid the hospital US\$930,000 since 2002 to grow,



Identity crisis: stem cells all look the same.

characterize and distribute Miz-1. But Roh told *Nature* that in December 2003, researchers at the hospital identified a chromosomal abnormality in that line. In April

2004 they switched to Miz-5, but pretended they were shipping Miz-1.

The subterfuge came to light during the police investigation into the research fraud committed by cloning expert Woo Suk Hwang (see *Nature* 439, 122; 2006). The human eggs used by Hwang were harvested at MizMedi. Roh admitted in November 2005 to paying for the eggs that he passed to Hwang, despite previously claiming they were from volunteers.

Roh says he knew nothing about the stem-cell switch, and only found out from the police report. But he says: "It's a healthy line. From a research perspective, it wouldn't

D. SCHARF/SPL



SCIENCE TEXTBOOKS OF THE FUTURE

Toss out the old backbreaker — it's time for video games.

www.nature.com/news



G8 leaders: will they be in the mood to listen to the world's scientists?

go to particular countries or projects, making it hard to organize a global plan that targets areas in greatest need.

But Capron says scientists can only be “deliberately optimistic” about such realities, and hope to influence decisions “through constantly repeating the same messages, and making the scientific communities’ positions

WHAT THE ACADEMIES WANT

Reinventing disease surveillance

- Efforts to coordinate disease surveillance across national and international agencies and research bodies
- Independent audit to recommend how to develop global surveillance
- Research into more rapid vaccine production methods
- Greater cooperation between human- and animal-

health communities

- Better collection and sharing of clinical and epidemiological data

Investing in energy R&D

- Highlight ‘reality and urgency’ of global energy supply
- Big, long-term infrastructure investments in cheap, clean, sustainable energies
- Boost developing countries’

capacity in innovative energy technologies

- Incentives to develop clean fossil, nuclear and renewable technologies
- Focus public research and technology efforts on energy efficiency, non-conventional hydrocarbons and clean coal, innovative nuclear power, distributed power systems, renewable energy sources, and biomass production

known”. Rees agrees: “For 12 academies of leading countries to emphasize the importance of these issues is an important signal to governments that they need to be addressed.”

Consensus building

So how does one go about getting scientists from 12 countries to agree on hot topics such as avian flu and energy? Capron says two to three members of each academy met a few months ago in Moscow to thrash out ideas and produce draft statements. They took these home for discussion, then after much e-mailing about wording, joint texts were agreed. There was a “spontaneous consensus” on the major steps needed for infectious diseases, according to Capron, with a “more protracted” discussion on energy.

All Capron is prepared to say about initial disagreements on energy is that the United States differed with Europe (France in partic-

ular) on the degree of support for research into nuclear energy, compared with other technologies such as carbon sequestration. France is highly dependent on nuclear energy, and lacks the coal reserves of the United States. But these were “divergences” rather than “aggressive differences”, he says diplomatically.

This kind of collaboration among the academies will improve with time, he adds, in particular by better involving academy members in the process. Lack of consultation marred the first joint statements on climate change last year, when the Royal Society issued a press release with its own interpretation of the consensus statement, leading to cries of foul play and spin by its partners. Such teething troubles won’t occur this year, assures Capron. He has now written rules on the procedures, which include requiring any press releases to be jointly agreed by all.

Declan Butler

cause any trouble.” According to Roh, around 80 Korean groups and more than 30 foreign groups were shipped the Miz-1 line. NIH spokesman John Burklow says three dozen US researchers received what they thought was Miz-1. But Roh says about 16 groups were shipped Miz-5 disguised as Miz-1.

The NIH has suspended research on MizMedi’s cell line, and insists no one used it in federally funded work. But the episode raises the question of what can be done to ensure the integrity of stem-cell lines.

James Battey, head of the NIH Stem Cell Task Force, says that when the agency receives a stem-cell line, it checks that the cells have normal chromosomes, are free of contamination, and replicate in

culture. These tests confirm a line’s health but do not check its identity. As the two MizMedi lines were both male, “the NIH would have no easy

“It’s a healthy line. From a research perspective, it wouldn’t cause trouble.”

way of knowing that Miz-5 was substituted for Miz-1”, he says.

Extensive DNA analysis would tell cell lines apart but is time-consuming, and reference information for the original line is often not available. Scientists at the Burnham Institute for Medical Research in La Jolla, California, working with a large group from the NIH and the DNA-analysis company Illumina of San Diego, have recently

made a dent in this problem.

The work was spearheaded by Mahendra Rao, who left the NIH in October 2005 and is now at the company Invitrogen. While at the NIH, Rao offered to perform detailed genetic profiles of all president-approved stem-cell lines for free. He says suppliers weren’t always eager to share their resources, but he has analysed the lines he could get agreements on, using bead-based microarrays developed by Illumina to compare single nucleotide polymorphisms (SNPs). Rao has published his analysis of 7 of 22 cell lines included in the NIH registry (www.biomedcentral.com/content/pdf/1471-213x-6-20.pdf), and says a paper in press covers ten more.

Rao says a quick, cheap profiling

method is essential: “You need to know what you are working with, because mix-ups are possible.” The NIH says it knows of no other cases of deliberate fraud, but there have been accidental switches. Rao says he uncovered a case in which a US company was shipping out wrongly identified cells after a technician accidentally mislabelled vials: “We were able to inform the few people who got the mislabelled line.”

SNP analysis also reveals the genetic basis of different cell types. The work heralds a breakthrough in that respect, says Burnham’s Evan Snyder: “Only now is the research evolving to the point where people are starting to profile these lines and compare them.”

David Cyranoski and Erika Check

ON THE RECORD

“There are a lot of testicles around.”

Chris Barratt of Birmingham Women's Hospital explains why stem cells from testicles could be a useful alternative to those from human embryos.

“I haven't done any conquering, per se.”

Tom Robinson, an accounting professor, considers his newly identified genetic link to Genghis Khan.

“Swabbing butts is not my highest priority, but it's a national emergency kind of thing.”

Biologist Rick Lanctot bites the bullet and samples shorebirds in Alaska for the H5N1 flu virus.

Sources: *The Guardian*, *Miami Herald*, *Reuters*

SCORECARD

**Sporting ants**

Scientists fiddle with ants' pheromones to whip up team rivalries in the world's tiniest football match: the Ant World Cup.

**Horse whispering**

An acoustics scientist finds that horse whinnies are far more complex than originally believed.

**Young footballers**

Kids playing football in new boots can end up with toxic shock syndrome from blisters, doctors report.

NUMBER CRUNCH

Have jet fuel, will travel. Inventor Brian Walker is building a massive, cross-bow-style rocket launcher in a bid to fire himself 32 kilometres in to the air this autumn.

6,000 newtons is the thrust Walker hopes his home-made rocket will achieve.

7 metres is the length of the carbon-fibre 'bow' Walker will use to launch the rocket.

\$15,000 is the cost of Walker's safety gear: a surplus Russian space suit.

Source: *Wired*

Cash for papers: putting a premium on publication

TOKYO

Financial rewards for publishing high-profile papers are spreading. Starting later this month, South Korean researchers will receive US\$3,000 from the government when they publish in elite journals. And that's a pittance compared with China, where some scientists can rake in more than ten times that amount.

As institutions and countries strive for international recognition, some are hoping that publication bonuses will help. But critics fear that this strategy could lead to a dangerous fixation on a few indicators of scientific success.

The Korean initiative, funded by the Ministry of Science and Technology, will award 3 million won (US\$3,000) to the first and the corresponding author on papers in key journals. “The plan is part of efforts to raise the morale of scientists who boost development in the country's science and technology sector,” says Young Nam Lim, deputy director at the ministry's department of basic science policy. A ten-member committee composed of ministry officials and researchers will choose the relevant journals. They are likely to include *Nature*, *Science* and *Cell* among others, Lim says.

Similar practices are already in place elsewhere. In Pakistan, under a system introduced by the science ministry in 2002, researchers can receive \$1,000 to \$20,000, based mainly on the cumulative one-year impact factor of their publications. Half is given as a research grant and the rest for personal use.

In China, bonuses are left up to the individual institution. China Agricultural University in Beijing, for example, will pay up to \$50,000 for high-impact papers, says its president Zhang-liang Chen. “This is not a big deal for great papers,” he says.

The Chinese Academy of Sciences' Institute of Biophysics in Beijing has a scale tuned to impact factors. Authors published in journals with an impact factor between 3 and 5 receive 2,000 yuan (\$250) per point, while a factor over 10 earns 7,000 yuan (\$875) per point. A paper in *Nature*, *Science* or *Cell* earns 250,000 yuan (\$31,000). The institute has had several such papers published over the past few years.

Advocates of the bonus schemes say that the incentives compensate for the low academic wages in the countries concerned. A bonus of

\$1,000, for instance, could be three times the monthly salary of a Pakistani university lecturer. Thanks partly to the incentives, the output of papers has increased dramatically, says Atta-ur-Rahman, chairman of the Pakistan government's higher education commission, who helped introduce the reward system.

But critics charge that more papers aren't necessarily better. “The bonus plan has had a devastating effect,” says Pervez Hoodbhoy, a physicist at Quaid-i-Azam University in Islamabad. “University researchers are rushing to publish by hook or by crook, and scientific and academic ethics are ignored in this haste.” He suspects that many papers with little significance are churned out, and says that bonuses could be adding to the problems of plagiarism and faked data that he has seen and heard.

“Good papers are the product of sweat, joy and sorrow. If this is calculated as 3 million won, I feel insulted.”

Others feel that the bonus system belittles their work. “Good papers are the product of sweat, joy and sorrow,” says Sunyoung Kim, a biologist at South Korea's Seoul National University. “If this is calculated to be 3 million won, I feel insulted.”

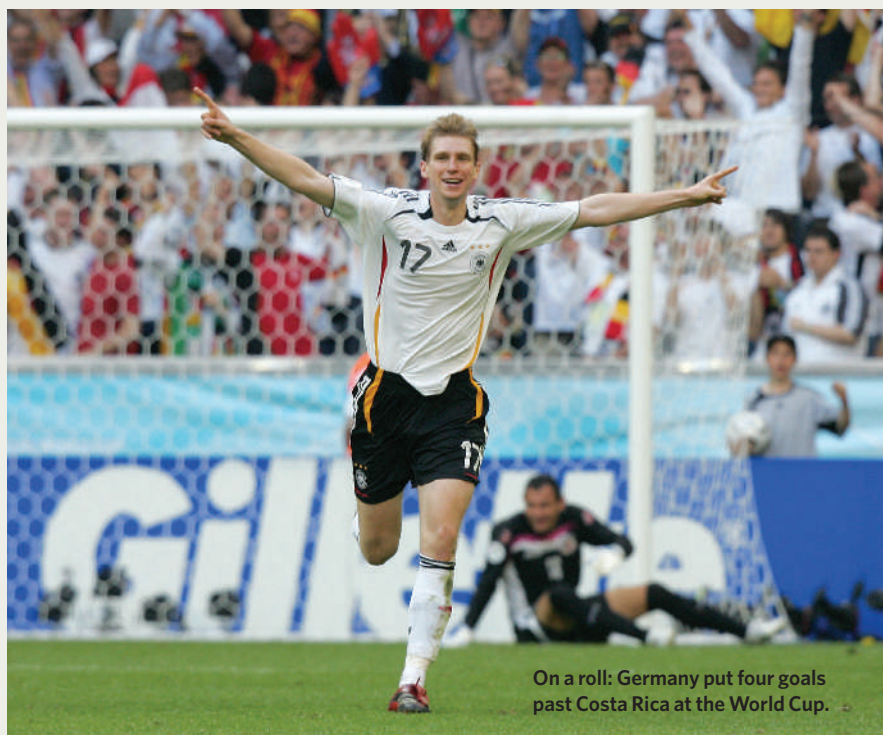
Yuan Tseh Lee, a Nobel prize-winning chemist, agrees. “There is too much pressure on scientists to get recognition in China,” says Lee, who is president of Taiwan's Academia Sinica. “If you just aim for fame and money, you will make yourself and your students miserable.”

Partly because of such concerns, the Shanghai Institute of Biological Sciences cancelled its bonus program in 2003. But Chen argues that at the moment bonuses are necessary, especially in China where a socialist system has made it difficult to reward hard work. “It's a necessary stage,” he says. “In ten years I'm sure the situation will be different.”

Peter Cotgreave, director of the London-based lobby group the Campaign for Science and Engineering in the UK, says reward systems have some merit, in part because the people effectively judging the award — anonymous peer reviewers — are independent of the award. Deserving researchers might miss out, he says, when, for example, high-quality papers are rejected by top journals. But he doubts that poor work would be rewarded, as researchers “only win if they are good enough to get into *Nature* or *Science*”.

Ichiko Fuyuno and David Cyranoski

P. HUGUEN/AFP/GETTY



On a roll: Germany put four goals past Costa Rica at the World Cup.

Goal fever at the World Cup

As the World Cup kicks off in Germany, mathematicians say they have proved one of soccer's classic clichés — the idea that when a team scores, the resulting confidence boost helps players romp to victory with a flurry of goals.

Although often seen as wishful thinking rather than football fact, such goal fever can be seen in football results stretching back decades, says mathematician Martin Weigel of Heriot-Watt University in Edinburgh, UK.

Much research has been geared towards improving tactics and understanding fan behaviour, but little has been done on factors affecting the distribution of final scores. Small studies carried out since the 1950s have generally assumed that a team has a set chance of scoring in a game — this depends on various factors and remains roughly constant throughout a match. This produces a 'normal' distribution of final scores,

like a symmetrical bell-curve.

But when Weigel and his colleagues at Germany's University of Leipzig examined results from the German men's and women's league (about 22,000 games) and from the finals of all previous World Cup tournaments (3,400 matches), they found that the scores didn't fit such a distribution: high-scoring games happened far too often.

This suggests that teams don't simply score a number of goals proportional to their skill, but rather are spurred on to greater heights after they score. The researchers designed a model in which a team's chances of scoring are multiplied for every goal, and found that it fitted the skewed distribution of final scores perfectly (E. Bittner, A. Nussbaumer, W. Janke & M. Weigel; preprint available at <http://arxiv.org/abs/physics/0606016>).

The effect is more noticeable in lower-standard

leagues than in the World Cup finals, probably because in higher-quality games teams are more evenly matched and so less likely to gain a psychological upper hand. But the boost gained from scoring is likely to be stronger in football, where goals are relatively rare, than in rugby or basketball, say, where both teams would expect to score reasonably often.

Similar effects can be found in other fields, however. Statisticians refer to situations where the outcome of one experiment influences the outcome of following ones as 'contagious distributions'. This has been seen with accidents, for example, where one mishap can make someone feel insecure and more likely to suffer further calamities.

Michael Hopkin

For more on the science of soccer, from goals to gambling, visit news@nature.com's interactive World Cup special at www.nature.com/news.



ANTIBIOTICS ABRIDGED
Unnecessarily long prescriptions may fuel drug resistance.
www.nature.com/news

Born or made? Debate on mouse eggs reignites

Experiments published this week fuel the controversy about whether mammals are born with a fixed pool of egg cells or can make fresh eggs. The issue is divisive not just because it challenges 50 years of scientific dogma. If women could generate new eggs it would have huge implications for fertility treatments, menopause, and even cell-based therapies.

Last year, a controversial paper¹ by Jonathan Tilly and his colleagues at Harvard Medical School proposed that egg cells can arise from bone marrow and other circulating cells². But Amy Wagers, also at Harvard Medical School, and her colleagues now report that this is not the case, at least for ovulated eggs³. Some critics think the new study proves Tilly's work was flawed, but Tilly and other experts say that some aspects of egg renewal remain open questions.

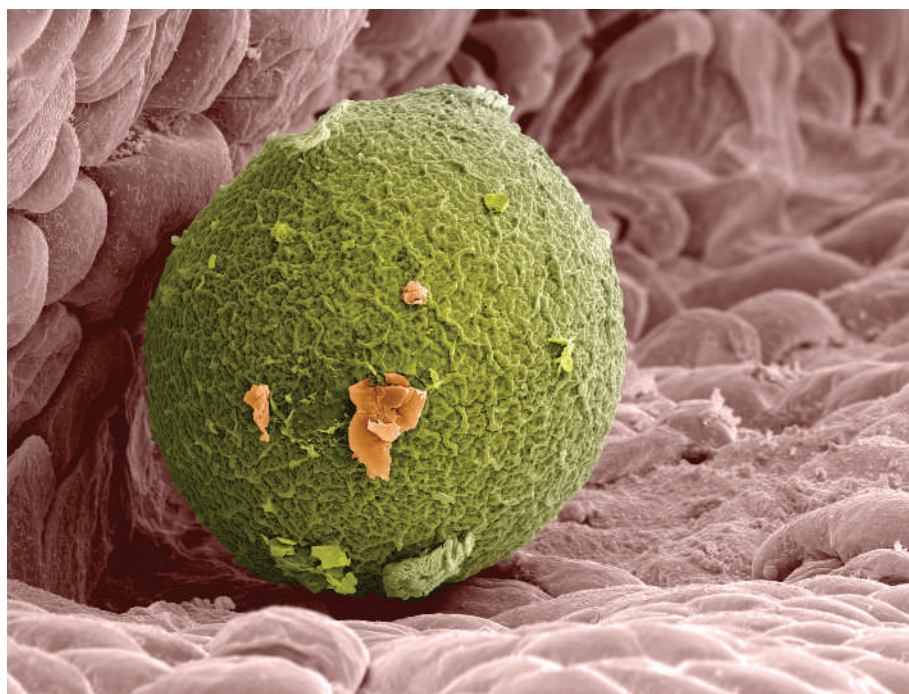
Tilly's bone-marrow study drew loud criticism from others in the field, both in the press and the literature. But without experimental testing of the ideas, much of the disapproval rang hollow. Now, Wagers and her colleagues have addressed the bone-marrow hypothesis. "This is a pretty powerful denial of the idea that new eggs form and contribute to fertility," says co-author Roger Gosden of Cornell University's Weill Medical College in New York.

Tilly's study found that a bone-marrow transplant from a donor mouse expressing a green fluorescent marker called GFP in all its cells resulted in green eggs — presumably newly generated — in the ovary of a normal mouse recipient.

To test that idea, Gosden and Wagers linked the circulatory systems of paired mice — one normal, and one with GFP. After several months, the researchers induced ovulation in both animals. Normal mice ovulated only normal eggs; GFP mice ovulated only green eggs, showing that there had been no exchange of egg-generating stem cells.

"It incontestably shows that the Tilly work was simply not true," says ovarian researcher David Albertini of the University of Kansas Medical Center in Kansas City. Echoing that sentiment, reproductive biologist Evelyn Telfer at University of Edinburgh, UK, is also unconvinced by Tilly's data, adding that the story changes as new data come in.

"Science is supposed to be a moving target," retorts Tilly. "Whether our data are repeatable is valid, but no one as yet has disproved any of our data." Tilly has several technical issues with



Seed of doubt: biologists are divided over whether adult mammals can regenerate egg cells.

the Wagers study, but he is most disturbed by the comparison of his work, which examined the pool of immature egg cells inside the ovary, to this study, which examines ovulated eggs outside the ovary.

He argues that circulating stem cells may still become immature egg cells inside the ovary. Such immature cells, he points out, are critical to fertility and hormonal balance, even if they are never ovulated. And he stands by the bone-marrow connection, pointing to additional experimental data from his group currently under review.

Embryonic idea

Albertini says this explanation is "looking to keep something of the original idea alive". He and other critics question why regeneration would occur in the immature but not the ovulated population of cells: "It's a bit of a stretch."

But several other experts agree that because Wagers's study did not examine immature cells inside the ovary, it leaves room for possible regeneration and the hunt for the responsible stem cells. But even if that were the case, all admit that it is a huge leap to suggest that the same happens in humans.

For now, regeneration in mice "is still an open issue", says Jeff Kerr of Monash University in Clayton, Australia. At least, "until someone comes up with some killer experiments to prove that bone marrow, or the ovary, is, or is not, the source of new oocytes". Kerr, Jock Findlay and colleagues add a piece of support for regeneration in research to be published online later this month⁴.

They found that the mouse strain used by Tilly and Wagers maintains a steady pool of oocyte numbers from before puberty into middle age, rather than the number declining with age as the dogma holds. Whether that is due to regeneration is unknown, says Kerr, but such a possibility would not be "completely out of left field".

Even Gosden, one of Tilly's more vocal critics, leaves the door open to the possibility of regenerative, albeit dormant, egg stem cells. "Our work suggests they do not contribute to new ovulations, but perhaps under special circumstances they could be brought back to life [in the lab dish]," he says. "I think the Tilly group might be right to some extent on that one."

Kendall Powell

1. Johnson, J. et al. *Cell* **122**, 303–315 (2005).
2. Ainsworth, C. *Nature* **436**, 609 (2005).
3. Eggan, K. et al. *Nature* doi:10.1038/nature04929 (2006).
4. Kerr, J. et al. *Reproduction* doi:10.1530/rep.1.01128 (2006).

CLOUDS HILL IMAGING/CORBIS

NUCLEAR REINCARNATION

Using nuclear power on a grand scale requires that spent nuclear fuel be reused. **Emma Marris** finds out which of the world's nations could jump on a reprocessing bandwagon.

It used to be that people were either pro or anti nuclear power — and that they would let you know which through a sticker on their car bumper. Now the debate is shifting. As evidence for global warming mounts, getting rid of nuclear power, with its very low carbon emissions, looks harder to justify in today's world than it did in the 1970s; hence the talk of a 'nuclear renaissance'. But there is a difference between keeping today's nuclear-power capacity (about 365 GW of capacity, which is responsible for generating roughly 16% of the world's electricity) and greatly increasing it. With extensive nuclear expansion depending on reprocessing spent nuclear fuel, whether or not to reprocess is now shaping up to be a new dividing line in the nuclear debate.

Reprocessing retrieves plutonium and unused uranium from used nuclear fuel. If you want to make sophisticated nuclear weapons, you need to have a reprocessing capacity, and the world's main nuclear-weapon powers all do. But not all of them have the capacity to reprocess spent fuel for civilian purposes. The United States has eschewed reprocessing as a way of making reactor fuel for the past 30 years. Now it is reconsidering it, sparking a reappraisal of the technology around the world.

Reprocessing makes usable fuel out of unusable waste, and in doing so reduces the volume and activity of what's left behind. The proposed conversion to civil reprocessing in the United States is being spurred in part by a desire to limit the amount of waste that needs to be put into long-term storage at the contentious Yucca Mountain site in Nevada. Supporters point to the reduction of waste and the increase in the amount of energy that can be extracted from a fixed amount of fuel as 'green' credentials for the technology. They also point out that the world's uranium supplies may not be sufficient to support an aggressive expansion of nuclear power unless fuel is reused.

But reprocessing is also accident-prone, expensive and makes available the sort of stuff that can be used to build bombs. No country has yet managed to make reprocessed fuel cheaper than the enriched uranium that is used in most reactors,



The United States is mulling over a proposal from the Bush administration to return to reprocessing, which the country has abjured due to proliferation concerns since India tested a bomb made from extracted plutonium in 1974. The new scheme, called the Global Nuclear Energy Partnership (GNEP), would have nuclear-weapon states plus Japan sending fuel to other states for use in their standard reactors. The donor countries would then take the fuel back for reprocessing. The idea has been received frostily in some quarters. The House of Representatives allocated about half of the funds requested by the administration to the project, saying there was too little detail in the plan to warrant a \$250-million investment. Many scientists and activists — both pro and anti nuclear power — have criticized the economics, timing and safety of the plan.

undercutting any economic rationale at today's uranium prices.

The reprocessing method the US Department of Energy proposes, called UREX+, aims to reduce the possibility of reprocessed plutonium being used in weapons by leaving it mixed with other highly radioactive metals. This supposedly leaves it too radioactive for malefactors to handle. But many opponents believe that the methods being discussed are still a security risk.

The issues are a little too complex to get on to a bumper sticker. But the current state of play can be displayed on a map.

2,400

The United Kingdom is home to the Sellafield site, where a reprocessing plant called Thorp has been inoperative since the discovery in April 2005 of a broken pipe that was leaking uranium and plutonium in a sealed-off cell. The incident renewed talk of shutting down the plant, which has been charged with polluting the Irish and North Seas. The publicly owned contractor that runs the whole Sellafield site, including two reprocessing plants and a hodgepodge of retired infrastructure needing clean-up, is up for sale.

400

Russia reprocesses its waste to form uranium, which it uses again as fuel, and plutonium, which it stores. Plans to build fast reactors that could burn pure plutonium have been edging forwards for years. Russia would like to import waste from other countries to stake a claim in a future 'plutonium economy', in part because it has little domestic uranium. According to Russian news reports, the head of the Mayak reprocessing plant, Vitaliy Sadovnikov, was charged with polluting the river Techa in March, but given amnesty in May because it was the 100th anniversary of the national state legislature.

Sweden is now thinking of reprocessing the waste from its nuclear-power industry for the first time in 20 years. A shipment of waste will set sail for Sellafield in the summer of 2007, if all goes as planned.

1,700

France is the king of reprocessing (the United Kingdom's Thorp plant never ran at capacity). Nuclear company Areva has plants at La Hague that have reprocessed waste from France's own extensive nuclear-power industry, as well as from Belgium, Switzerland, the Netherlands and Japan. In 2004 and 2005, weapons-grade plutonium that the United States promised Russia it would get rid of was reprocessed at Areva's Cadarache fuel plant into mixed-oxide (MOX) fuel. The MOX fuel was then used at the Catawba Nuclear Station in South Carolina.

40

Japan has just begun a 17-month test phase at the brand-new Rokkasho reprocessing plant, which is far larger than the Tokai pilot reprocessing plant, now headed for retirement. The new plant's capacity will be 800 tonnes per year. Japan has also agreed, on paper, to join the GNEP, but a 1988 agreement banning the transfer of reprocessing technology between Japan and the United States is standing in the way.

275

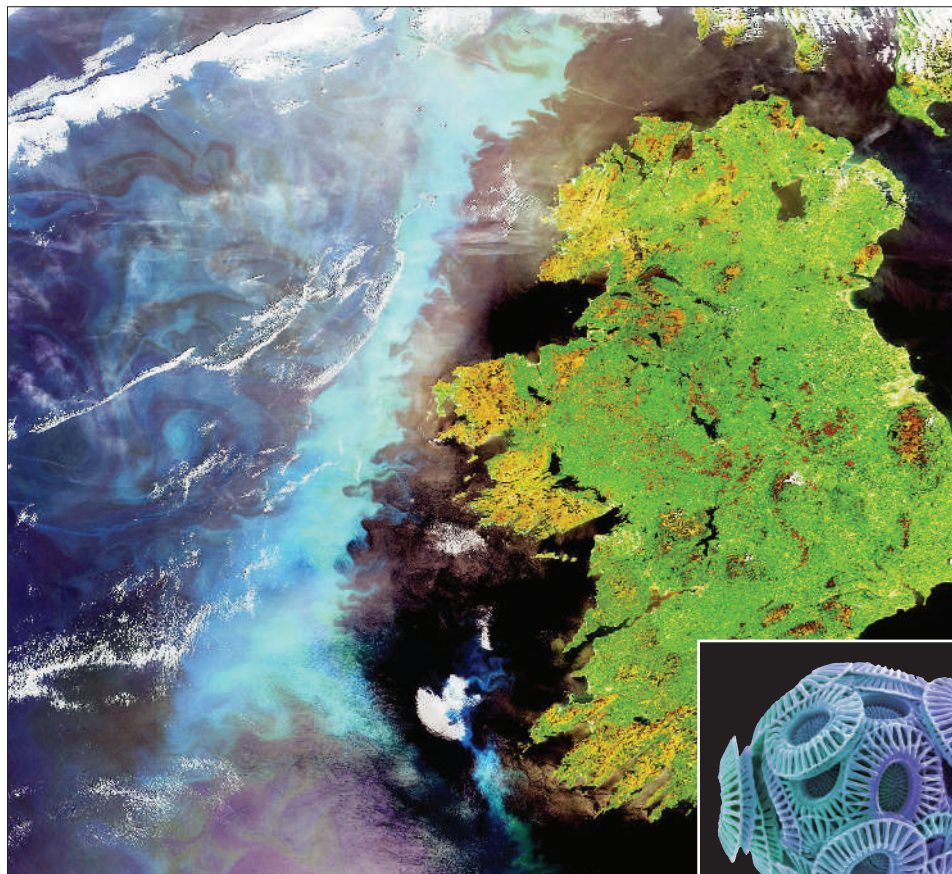
India reprocesses at three small plants. Famously, it got the technology from the United States, then used the plutonium it extracted to make a nuclear bomb, which it tested in 1974. India was affronted when asked to join the GNEP not as a provider of reprocessing power but as a client.

Australia's prime minister, John Howard, announced on 6 June that the nuclear-free country will study the possibility of building some nuclear power plants, and of becoming a reprocessing nation. Howard visited US energy secretary Sam Bodman last month to discuss the GNEP reprocessing scheme.

- Countries with commercial reprocessing capability
- Countries with nuclear power
- Countries with neither



Reprocessing capacity
(tonnes per year)



SNAPSHOT Atlantic in bloom

The death throes of a giant plankton bloom (the milky blue ribbon in the photograph) have been captured by Envisat, an earth-observation satellite operated by the European Space Agency. The algal bloom, snapped on 6 June, is among the largest recorded — at around 500 km long, it stretches the full length of Ireland.

The milky water, which would appear chalky-white to anyone on a boat sailing through it, forms as the plankton, having used up all the available nutrients in the sea around them, start to shed their calcite shells. The species involved in this particular display is likely to be *Emiliania huxleyi* (inset), which often forms blooms in the North Atlantic at this time of year.

Photosynthesizing plankton species such as *E. huxleyi* are at the root of almost all food webs in the ocean, being eaten by animal plankton. But despite their visual impact, little is known about how the blooms affect the local ecosystem.

Jim Giles

ESA, S. GSCHMEISSNER/SPL

US satellite system loses climate sensors

WASHINGTON DC

Climate scientists are scrambling to find new backers for atmospheric sensors they had hoped would fly on board a US military-civilian polar satellite system. Last week, government officials scrubbed three planned instruments — which would have measured key climate variables — in an effort to get the long-troubled project back on budget and schedule.

First proposed in 1994, the National Polar-orbiting Operational Environmental Satellite System (NPOESS) had been touted as a one-stop shop for both military and civilian needs to gather data on global weather. It was meant to save money and avoid duplication between three government agencies — the US Department of Defense, the National Oceanic and Atmospheric Administration (NOAA), and NASA. Originally slated to cost US\$6.5 billion and launch in 2008, the project is going to cost upwards of US\$11 billion and will not launch until at least 2013.

Under the new design, the programme “will have fewer satellites and less sensors, while costing more money”, NOAA administrator Conrad Lautenbacher told a congressional hearing on 8 June.

The cutbacks come as part of a defence department review of the programme, required by law because it was so far over budget. Five key sensors have been temporarily removed, three of which are related to climate science.

“They decided to de-scope everything not related to short-term weather forecasting,” says programme scientist P. K. Bhartia, who works on an ozone sensor for the satellites. “At this point, we’re really confused as to what will happen.”

If industry, a university or a research agency decides to pay for the instruments directly, they — or stripped-down versions of them — may be placed back onto the NPOESS satellites. In the meantime, climate scientists are left wondering whether their sensors will ever get to fly.

“Absolutely, we’re quite disappointed,” says Ron Isaacs, executive vice-president of the Massachusetts-based company Atmospheric Environmental Research, which helped design another sensor that was axed. That instrument, Isaacs says, would have served some unique functions, such as measuring soil moisture.

The European Meteosat satellites could work with NPOESS satellites to obtain some of the necessary data. But US officials are concerned about how much of the data would actually be shared — especially information useful for the military.

Still, many in Congress support the NPOESS programme, even as it balloons in cost. “We have to make this work,” says congressman Sherwood Boehlert (Republican, New York), chairman of the House Committee on Science. “At some point that word ‘operational’ has to mean more than adding a vowel to the acronym.”

Jacqueline Ruttimann

Computations cast doubt on chemical synthesis

An impressive 37-step chemical synthesis published this year (J. La Clair *Angew. Chem. Int. Ed.* 45, 2769–2773; 2006) has been called into question.

Scott Rychnovsky, a chemist at the University of California, Irvine, says the synthesis of the chemical hexacyclinol is “an impossibility” if his computational paper is right (S. D. Rychnovsky *Org. Lett.* doi:10.1021/ol0611346; 2006).

Rychnovsky's computer predictions suggest that the molecule was assigned the wrong structure when first isolated from a fungus in 2002. So what was synthesized? The author of the *Angewandte Chemie* paper, James La Clair of the Xenobe Research Institute in San Diego, California, says: “I didn't make it up. Maybe I synthesized a compound that is not hexacyclinol but something else, and it needs a new name.”

La Clair has been engaging with critics on at least one chemistry blog. An editor at *Angewandte Chemie* says the journal will look into the matter.

Historic observatory to be preserved in luxury complex

Spa-goers and star-gazers can live side by side at the Yerkes Observatory. The University of Chicago has sold the 109-year-old building and surrounding site to a land developer for roughly US\$8 million.

The New York-based Mirbeau Company plans to build a 100-room spa and 72 luxury homes on part of the site, pending approval by the nearby village of Williams Bay, Wisconsin. The observatory and most of the remaining land will be preserved as an education and outreach facility — run by a non-profit organization and funded in part by room and property taxes from the new buildings.

Antiquated and hampered by light-



P. BASSETT/SPL

Changing world: a spa may soon be built next to the once legendary Yerkes Observatory.

polluted skies, the observatory has had limited use as a research facility since the 1970s. Among those who used its legendary 40-inch telescope are astronomer Edwin Hubble and astrophysicist Subrahmanyan Chandrasekhar.

Geologists find clues to Chinese quake prediction

The background to one of the most famous earthquake predictions in history has now been established, after a team of scientists gathered declassified documents and interviewed local people to determine what really happened.

In February 1975, local governments ordered some areas around Haicheng, a town in northeast China, to be evacuated. A magnitude-7.3 quake struck, and Chinese officials claimed that the prediction had saved thousands of lives.

But earthquake forecasting has enjoyed little success since. So Kelin Wang of the Geological Survey of Canada recently teamed up with three colleagues in China to document the history of the Haicheng prediction (K. Wang, Q.-F. Chen, S. Sun and A. Wang *Bull. Seismol. Soc. Am.* 96, 757–795; 2006).

They tracked down and translated documents, from secret reports to notes in geologists' logbooks. The region had previously been identified as at risk, but Wang's team found that local officials gave

Cloud satellite pierces the gloom



NASA/JPL/CIRA

Earth's clouds are getting their best-ever internal scans, now that NASA's CloudSat satellite is up and running. This image, of a warm-front storm over the Norwegian Sea on 20 May, is the first returned by the spacecraft.

The colours represent particles of different reflectivities, with red indicating highly

reflective particles such as water droplets or ice crystals. Blue indicates thinner, cirrus-type clouds.

CloudSat and its partner satellite CALIPSO are building up a global picture of atmospheric water droplets and aerosols, which is designed to resolve the three-dimensional structure of clouds (see *Nature* 437, 468–469; 2005).

orders to evacuate after a series of strong foreshocks. Unfortunately, foreshocks don't occur before every quake.

House votes against budget boost for the NIH

Call it the case of the disappearing cash. On 7 June, a subcommittee of the US House of Representatives voted not to raise the budget of the National Institutes of Health (NIH) next year. The vote would leave the NIH budget at \$28.3 billion — still making it the richest US science agency, but with an essentially flat budget after being cut last year for the first time in three decades.

Research advocates say this defies the expectations set in May, when the House voted to spend an additional \$7 billion on health and education programmes. "To see the NIH is getting nothing of that is disappointing," says Jon Retzlaff, director of legislative relations at the Federation of American Societies for Experimental Biology.

Advocates are pushing lawmakers to find more money for the NIH, but there is little time: the House hopes to finalize all budgets by the 4 July recess. The budget approval process is taking longer in the Senate, where lawmakers have promised to boost budgets.

German biodiversity study goes back to the land

A large-scale field study in Germany aims to investigate the relationship between land-use intensity and biodiversity in unprecedented detail.

The three-year, €8-million (US\$10-million) project funded by the DFG, Germany's main research-funding body, will be located at three test sites in different parts of the country. Each of the 100-square-kilometre 'exploratories' is divided into 1,000 observation units.

Researchers will use the field data to test ideas about how biodiversity changes as woodlands and grasslands change from semi-natural to intensively used environments. One goal is to gather reference data to be able to study biodiversity changes in the future in a more coordinated manner.

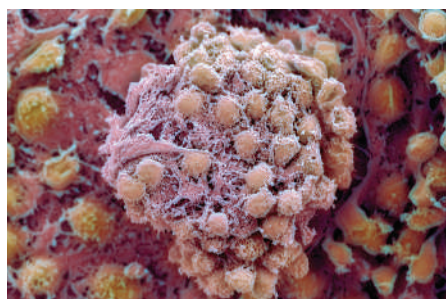
Brighter outlook for stem-cell research in Europe

International societies for stem-cell research have written an open letter of support to Italy's new research minister, Fabio Mussi, who has said he will withdraw Italy from a pledge to oppose the use of European Union (EU) money for human embryonic stem-cell research.

The EU's seventh Framework research programme, now going through the final stages of political approval, has provisions for such research — even though it is forbidden in some EU countries.

Six EU countries opposed to embryonic stem-cell research signed the pledge in November. Their combined weight would have blocked approval of the Framework programme provisions, but if Italy withdraws, the block would be lifted.

The letter, on behalf of the International Society for Stem Cell Research and the European Consortium for Stem Cell Research, applauds what it calls an "honourable decision" that would ensure that "Italy is no longer blocking scientific progress for universal benefit".



M. STOJKOVIC/SPL

Italy's *volte face* on embryonic stem-cell research could undo a previous EU block on member funding.

US universities cleared to continue stem-cell research

Two US universities are back in the race to grow embryonic stem cells matched to human patients.

Both the University of California, San Francisco, and Harvard University have begun work in the field that was set back when Woo Suk Hwang, South Korea's cloning pioneer, was found to have fabricated some of his published findings.

The work is highly controversial because human embryos are discarded in the process, and US researchers are banned from using federal research money for such studies. Harvard officials announced on 6 June that they had given the go-ahead after two years of regulatory review by eight committees. The teams plan to ask women to donate their eggs, but do not know whether they will undergo the invasive procedure for research purposes only.

The US teams now join some half-a-dozen groups around the world already working to grow human embryonic stem cells from cloned human embryos.

Correction

In the News Feature "An easy way out?" (*Nature* **441**, 570–571; 2006), B. Taylor Bennett should have been located at the University of Illinois at Chicago, not at the University of Chicago.

The tipping point of the iceberg

Could climate change run away with itself?

Gabrielle Walker looks at the balance of evidence.

"Be worried," *Time* magazine advised those looking at its 3 April cover showing a forlorn polar bear surrounded by puddles of melted ice. "Be very worried." Immediately below came the occasion for such alarming counsel: "Earth at the tipping point."

The idea that passing some hidden threshold will drastically worsen man-made climate change has been around for decades, normally couched in technical terms such as 'nonlinearity', 'positive feedback' and 'hysteresis'. Now it has gained new prominence under a new name. In 2004, 45 newspaper articles mentioned a 'tipping point' in connection with climate change; in the first five months of this year, 234 such articles were published. "Warming hits tipping point," one UK newspaper recently warned on its front page; "Climate nears point of no return," asserted another. The idea is spreading like a contagion.

The infectious analogy is appropriate. When the writer Malcolm Gladwell unleashed the idea of tipping points on the popular imagination in his book of the same name¹, he was comparing the way aspects of life suddenly shift from obscurity to ubiquity to effects normally studied in epidemiology. Gladwell's tipping points were manifestations of the catchiness of behaviours and ideas. The notion that climate change is getting out of control is catchy, and it has caught on in academic papers and political debates as well as headlines. But is the climate really on the point of tipping over into a radically different state? And if so, what are the implications?

A tipping point usually means the moment at which internal dynamics start to propel a change previously driven by external forces. The idea raises two questions. First, when will that moment be reached? Second, after it has been passed, is the system now destined to run its course regardless of what goes on elsewhere — is a tipping point a point of no return?

Although there's no strong evidence that the climate as a whole has a point beyond which it switches neatly into a

new pattern, individual parts of the system could be in danger of changing state quickly, and perhaps irretrievably. And perhaps the most striking of these vulnerable components are in the Arctic. Farthest north is the carapace of sea ice over the Arctic Ocean. South of that is the vast ice sheet that covers Greenland. And then there is the ocean conveyor belt, which originates in a small region of the Nordic seas and carries heat and salt around the world.

On thin ice

All three seem to have inbuilt danger zones that may deserve to be called tipping points. And the outside forces pushing them towards those points are gathering. "There is near-universal agreement that we are now seeing a greenhouse effect in the Arctic," says Mark Serreze from the US National Snow and Ice Data Center in Boulder, Colorado.

Serreze studies sea ice, the member of the arctic triumvirate that has had most recent attention. In the winter, sea ice more or less covers the Arctic Ocean basin. Summer sun nibbles at the pack ice, shrinking it at the edges and creating patches of open water within. Open water reflects much less sunlight than ice — it has what is known as a lower albedo — so the greater the area of dark open water, the more summer warmth the ocean stores. More stored heat means thinner ice in the next winter, which is more vulnerable to melting the next summer — meaning yet more warmth being stored in the open water in the following year, a cycle known as the 'ice-albedo feedback'. "Once you start melting and receding, you can't go back," says Serreze.

It seems that some of this process is under way. Serreze and his colleagues have found that the summer sea ice has shrunk by an average of 8% a decade over the past thirty years². The past four years have seen record lows



"There is near-universal agreement that we are now seeing a greenhouse effect in the Arctic."
— Mark Serreze



Pointing tip: an iceberg off the coast of Greenland, shed from the edge of the island's icepack.

in the extent of September sea ice, and in 2005 there was 20% less ice cover than the 1979–2000 average, a loss of about 1.3 million square kilometres, which is more than the area of France, Germany and the United Kingdom combined. It was this finding that triggered a raft of alarming headlines.

The ice's volume, rather than its extent, would be a more useful figure, but this is hard to estimate. Radar measurements showing how proud the ice sits with respect to nearby water would help, but the European Cryosat mission intended to provide these data was lost on launch in October 2005. A refligh is planned, but at present the only way to determine the pack thickness is from below. In 2003 Andrew Rothrock and Jinlun Zhang of the University of Washington in Seattle analysed results from a series of submarine cruises from 1987–97 and concluded that the ice thinned by about one metre during that period³.

Flaming January

A natural swing in wind and weather known as the Arctic Oscillation may have played a key role in the decline. In 1989, this index began to approach its positive mode, in which a ring of strong winds circles the pole. Zhang and his colleague Roger Lindsay, also at the University of Washington, believe these winds flushed large amounts of thick ice out of the Arctic through the Fram Strait, east of Greenland. Last year, they published a model suggesting that because the replacement ice was thinner and more vulnerable to the ice–albedo feedback, this extra loss pushed the Arctic over the edge. Their paper's title: "The thinning of Arctic Sea Ice, 1988–2003: Have We Passed a Tipping Point?"⁴.

But given that sea ice was disappearing even before the Arctic Oscillation lurched into its positive state, it is unlikely to have been the sole trigger. "The Arctic Oscillation was a strong kick in the pants," says Serreze, "but if we hadn't had it we would still have seen the ice loss."

Whatever the precise mechanisms, the decrease in ice seems to be warming the atmosphere, as heat pours from

the open water into the air above it. Springtime temperatures began rising throughout the Arctic basin in the 1990s⁵. This year, the Arctic archipelago of Svalbard experienced a remarkable heatwave. January was warmer than any previously recorded April, and April was more than 12 °C warmer than the long-term average.

Lindsay and Zhang suggest that the ice–albedo effect has indeed passed a tipping point, with the internal dynamics more important than external factors. But neither observations nor models suggest that the effect will now run away without outside help. According to climate modeller Jason Lowe of the UK Met Office in Exeter, the relationship between sea ice and temperature is reassuringly linear. "When you plot sea ice against temperature rise, whether from observations or models, it forms a remarkably straight line," he says. "It's not a runaway effect over the sorts of temperature ranges that we're predicting here." Lowe says that although the planet will almost certainly lose more ice, it does not have to lose it all. But if current trends in greenhouse-gas emissions and global warming continue, a planet that used to have two permanent polar caps will have only one.

Losing the sea ice would be bad news not only for polar bears and other charismatic megafauna, but also for some of the Arctic's smaller inhabitants. Photosynthetic plankton that live in pores and channels within the ice are the foundation of the area's food supply, and are not well adapted to ice-free life. Open-ocean plankton might benefit, but the Arctic is so poor in nutrients that this would probably not be much compensation⁶.

Change of winds

Compared with the overall scale of human-induced climate change, the additional warming expected if the ice–albedo feedback goes all the way would not be immense. The 4.5% of the Earth's surface above the Arctic Circle is simply too small to make a radical difference to the planet's energy balance. There are, however, some hints that the loss of sea ice may have more far-reaching effects beyond the simple number of watts absorbed per square metre. Tim Lenton, an Earth-systems scientist at the University of East Anglia in Norwich, UK, points out that our current, relatively stable pattern of winds, which is caused by three circulatory air systems in each hemisphere, depends in part on a white and cold North Pole.

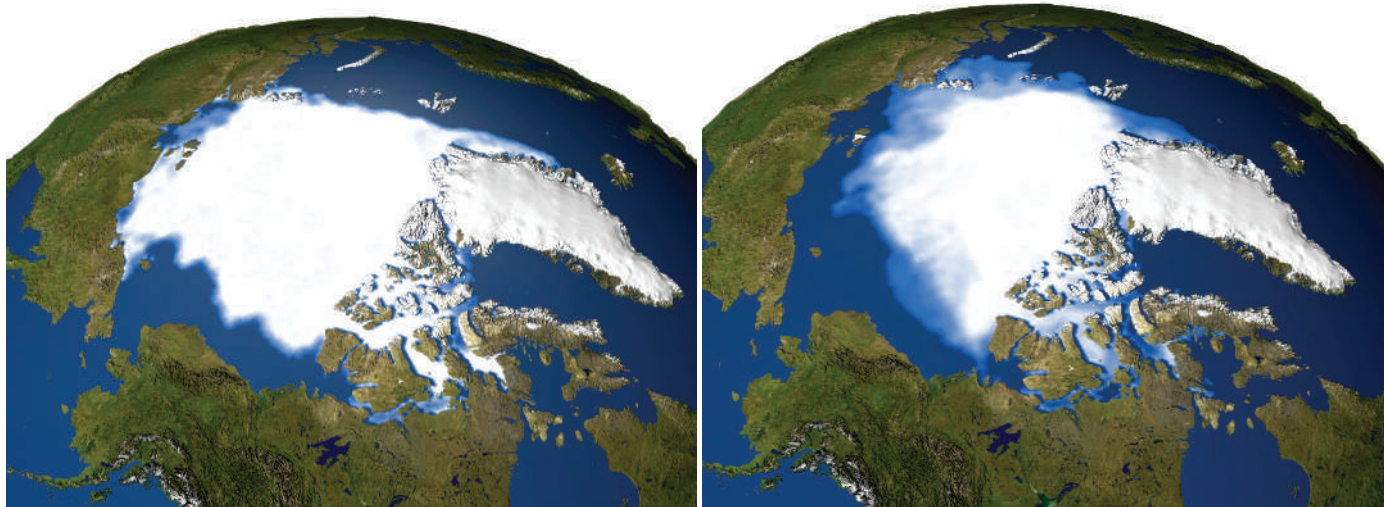
Sinking air in the Arctic is an integral part of an air system called a Hadley cell; there is another Hadley cell over the tropics. Between these two cells are the fierce westerlies and the high-altitude jet streams that drive storms around the middle latitudes. "If any part of the current structure broke down, that would be profound," says Lenton. "If the system starts to switch seasonally between three cells and a less stable structure, you change the position of the jet streams, you change everything." Models of this possibility

are scarce, but Jacob Sewall and Lisa Sloan of the University of California, Santa Cruz, have shown that an ice-free Arctic could shift winter storm tracks over North America, drying the American west⁷.

The local warming caused by less sea ice could also affect the second tipping point, the size of the Greenland ice sheet. Here the effects could be dramatic, although delayed by centuries; there is enough ice on Greenland to raise sea levels by seven



"If any part of the current structure broke down, that would be profound."
— Tim Lenton



metres. "After hurricane Katrina, the deepest water in New Orleans was six metres," says glaciologist Richard Alley from Penn State University. "Greenland is more than that for all the coasts of the world. Do you move cities, do you build seven-metre walls and hope they stay, or what?"

Until recently, nobody had painted a convincing portrait of how Greenland is responding to Arctic warming. A glacier here may recede while one over there grows; ice may be accumulating inland and eroding near the coast. But in the past couple of years, almost all of the indicators have started to point in the same direction. Greenland is melting.

Cracking up

Although satellite measurements of Greenland's interior suggest that snow has recently been accumulating there, the margins are receding⁸. Laser measurements taken from planes suggest that this coastal melting is probably enough to outweigh the build-up of snow inland⁹. Also, Greenland's glaciers seem to have been speeding up. A few months ago, Eric Rignot of NASA's Jet Propulsion Laboratory in Pasadena and Pannir Kanagaratnam of the University of Kansas, Lawrence, published satellite evidence that between 1996 and 2000, Greenland's more southerly glaciers had begun to accelerate, and that by 2005 the northerly ones had followed suit¹⁰. They estimate that over the past decade this lurching has more than doubled Greenland's annual loss of ice, from 90 to 220 cubic kilometres per year.

"In the past decade there has been a lot of warming," says Alley. "There's plenty of room to argue whether that's a natural fluctuation or not, but there's a clear relation between Greenland getting warmer and Greenland getting smaller."

Modelling by Jonathan Gregory from the University of Reading and his colleagues suggests that it would require an average warming worldwide of 3.1 °C to drive this shrinking to its ultimate conclusion of an ice-free Greenland¹¹. This climatic point of no return is around the middle of the range foreseen by the Intergovernmental Panel on Climate Change, but is higher than a previous estimate made by the same group¹². Their revision is a measure of how quickly the field is changing. "It's not just Greenland that is going fast," says Alley. "The rate of publications, the rate of new papers, and the rate of disagreement have multiplied amazingly."

But these models do not take into account the dynamism of Greenland's glaciers. In 2002 Jay Zwally from NASA's Goddard Space Flight Center in Greenbelt, Maryland, found that as soon as summer meltwater appeared on the surface of west-central Greenland, the ice began to slip more quickly¹³. This is surprising, as slip rates should

Thin on top: between 1979 (left) and 2005, the minimum extent of Arctic sea ice shrank by about 20%.

depend on processes at the base of the ice rather than at its surface. But Zwally points out that the great lakes of water produced by the melting could slip down conduits in the ice and be delivered directly to the bed.

This result doesn't necessarily make a big difference to the fate of Greenland, as the increase in the ice's speed was relatively small. But it points to a new way in which the ice sheet could react to climate change quicker than anyone had realized. "In places inland where the ice is frozen to its bedrock, if you warm the surface and wait for heat to get conducted to the bottom it takes 10,000 years," says Alley. "But if you send water down through a crack it takes maybe 10 minutes, maybe 10 seconds." If this process started to move inland, even the interior of Greenland's ice sheet could be vulnerable to warmer air. That could point to the sort of self-sustaining feedback that tipping points are made of.

Getting fresh

The models don't incorporate this mechanism, because they can't. The cliff fronts of many Greenland glaciers are shot through with bright blue conduits, but nobody knows how widespread these veins are inside the ice. Still, the responsiveness of Greenland's glaciers makes that point-of-no-return figure of 3.1 °C even less comforting. What's more, a lot of damage can be done without losing all of the ice. The ice sheet did not vanish during the last interglacial, around 130,000 years ago, when temperatures in the north were a few degrees higher than they are today. And yet the latest analyses suggest that meltwater from Greenland increased the sea level by between two and three metres. The only good thing about such an increase is that it would take centuries.

As with the melting of Arctic sea ice, the melting of the Greenland ice sheet has implications for its neighbours. The third tipping point is the origin of the great oceanic conveyor belt, or thermohaline circulation. Thanks to its cold temperatures and high salinity, water in the Nordic seas

between Greenland and Scandinavia is unusually dense and sinks. Surface water is drawn northwards to replenish this. One result of this flow is that Britain is warmer than its latitude would seem to deserve.

The sinking process sets a global mass of water in motion, transporting vast amounts of heat around the oceans. In the 1980s, models began to suggest that melting ice in the north could weaken this system, by putting a plug of fresh water over the



"The rate of publications, the rate of new papers, and the rate of disagreement have multiplied amazingly."
— Richard Alley

sinkhole. This led to fears of abrupt climate change and snap ice ages in Europe and eastern America. These days most scientists think that the power of this flow to affect European temperatures under current conditions, or in a globally warmed future, has been overestimated¹⁴. But changes in the system could still have far-reaching implications. And models suggest that the thermohaline circulation has its own tipping point.

Comparing the output from 11 different ocean and climate models, ocean modeller Stefan Rahmstorf from the Potsdam Institute for Climate Impact Research (PIK), Germany, has concluded that it would take between 100,000 and 200,000 cubic metres of fresh water per second to shut down the thermohaline circulation — similar to the outflow from the Amazon River¹⁵. And once the circulation is stopped, restarting it would take a lot more cooling than just reversing the system into the conditions in which it was previously working.

Dry Asia

The good news is that although the Arctic does seem to be getting fresher, it is nowhere near the danger point. Add together the increased output from disappearing sea ice (which moves fresh water from the point where sea water freezes to the point where the ice melts), the melting of Greenland and increased Arctic river flow and you still have barely a quarter of the lower bound of the model threshold.

However, measurements of flow in the deep ocean suggest that the circulation might be fluctuating in ways not considered by the models¹⁶. And if the melting of Greenland were to gather pace, the thermohaline circulation would be vulnerable. If the lower bounds of the models turn out to be right, a rate of melting that would get through the ice in 1,000 years would trouble the ocean overturning in centuries. “The fate of the thermohaline circulation will be decided by Greenland,” says Rahmstorf. “If that goes quickly it will be bad news for the deep-water formation. But if Greenland is stable, the risk of shutting down the circulation completely is very small.”

Any such shutdown would probably have only a small effect on European temperatures. But thanks to the Coriolis effect, says Rahmstorf, such a large shift in the ocean circulation would redistribute sea water so that the North Atlantic rose by up to a metre¹⁷. There are also suggestions that Atlantic fisheries could collapse.

But the biggest danger would come farther south. In the past, similar changes in ocean circulation seem to have led to significant shifts in tropical rainfall. “If you switch off the thermohaline circulation, the tropical rainfall belts shift. All the models show this. It’s quite simple robust physics,” says Rahmstorf. General circulation models, which try to simulate the workings of the climate system as a whole, often including the ocean, predict at least some weakening of the thermohaline circulation by the end of the century, with a knock-on effect on tropical rainfall — the system that provides much of Asia with food. And as with Greenland, the change doesn’t have to be complete to have consequences. “Just weakening the system is by no means harmless,” says Rahmstorf. “You’d get the same pattern of effects as for a total shutdown, but just a smaller amplitude.”

What insights do these three examples provide into tipping points more generally? One is that they are hard to

predict, because they often depend on phenomena too subtle or small to be captured in climate models — the effect of wind patterns on sea ice, or the flow of water through ice-sheet cracks. And bear in mind that the members of the Arctic triumvirate are in principle pretty simple, in that they depend for the most part on physics alone. Possible tipping points in which biology too plays a role — for example, the potential die-back of the Amazon forests if their area diminishes beyond a certain threshold — are even harder to get a grip on. As a result, few researchers are prepared to put a number on how much warmer we could allow the climate to get without endangering humans.

Human factors

Another insight is that points of no return may not be particularly important. Large-scale melting in Greenland is a serious issue over centuries regardless of whether it goes all the way. And the question of whether sea ice would continue to shrink without global warming is academic. The current level of greenhouse gases ensures that the world will continue to warm over the next decades, and the current structure of the world economy ensures that, over that time, there will be further increases in the greenhouse-gas level. The question is what will be done about it, and how soon. As Alley says: “The human tipping points are probably more important than the natural ones. It’s at what point the situation becomes intolerable to us that matters.”

The message from the Arctic is that there is still time to avert the worst potential consequences of the nonlinearities in our climate system. Although it is probably too late to stop a serious decline in sea ice, the other two more powerful members of the Arctic trinity look to be some way off their danger points. And unlike the thermohaline circulation, and perhaps the Greenland ice sheet, the change in sea ice does seem to be reversible.

If disappearing ice and dying polar bears can tip public opinion over its ‘be very worried’ threshold into the realms of greater action, then further tipping points in the human world might have their own, positive, role to play. Ottmar Edenhofer, an economist at PIK, found that in some economic models of responses to climate change, increasing carbon prices encourages renewable energy. Above a specific threshold, even if the price of carbon drops, the advantages of renewables have become irrevocable and the move away from fossil fuels continues¹⁸. “Our task is to find a way to kick the economic system into a new equilibrium and we can use the tipping points of the market to achieve that,” says Edenhofer. “Tipping points are part of the problem, but they could also be part of the solution.” ■

Gabrielle Walker is a writer who specializes in Earth science and cold places.

“The fate of the ocean circulation will be decided by Greenland. If that goes quickly it will be bad news.”
Stefan Rahmstorf



- Gladwell, M. *The Tipping Point: How Little Things Can Make a Big Difference* (Little, Brown, Boston, 2000).
- Stroeve, J. C. et al. *Geophys. Res. Lett.* **32**, L04501 (2005).
- Rothrock, D. A. et al. *J. Geophys. Res.* **108**, 3083 (2003).
- Lindsay, R. & Zhang, J. J. *Clim.* **18**, 4879–4894 (2005).
- Overland, J. E. et al. *J. Clim.* **17**, 3263–3282 (2004).
- Smetacek, V. & Nicol, S. *Nature* **437**, 362–368 (2005).
- Sewall, J. O. & Sloan, L. C. *Geophys. Res. Lett.* **31**, L06209 (2004).
- Johannessen, O. M. et al. *Science* **310**, 1013–1016 (2005).
- Krabill, W. et al. *Geophys. Res. Lett.* **31**, L24402 (2004).
- Rignot, E. & Kanagaratnam, P. *Science* **311**, 986–990 (2006).
- Gregory, J. M. & Huybrechts, P. *Phil Trans. R. Soc. Lond. A* (in the press).
- Gregory, J. M. et al. *Nature* **428**, 616 (2004).
- Zwally, H. J. et al. *Science* **297**, 218–222 (2002).
- Schiermeier, Q. *Nature* **439**, 256–260 (2006).
- Rahmstorf, S. et al. *Geophys. Res. Lett.* **32**, L23605 (2005).
- Bryden, H. L., Longworth, H. R. & Cunningham, S. A. *Nature* **438**, 655–657 (2005).
- Levermann, A. et al. *Clim. Dyn.* **24**, 347–354 (2005).
- Edenhofer, O. et al. *Energy J.* **27** special issue, 57–107 (2006).



Eat your cake and have it

Reducing your calorie intake makes you live longer — if you're a rat or a worm. **Laura Spinney** asks whether the same holds for humans — and if it does, whether the benefits could be put in a pill.

It may be just a dot in the Pacific Ocean, but the Japanese island of Okinawa has long held a fascination for researchers into ageing. Its inhabitants include an unusually high number of centenarians, and fewer of its people die from diseases of old age such as cancer and stroke. Some have suggested the islanders' secret is their frugal diet. Now scientists are attempting to put this idea to the test — and are even developing drugs that they hope will produce the same effect. Could the path to a long and healthy life really be as easy as swallowing a tablet?

The issue divides researchers. Calorie restriction, or cutting energy intake below energy expenditure, can slow ageing, reduce mortality, and extend maximum lifespan in rats, mice, fish, flies, worms and yeast. But to date, the only evidence that it works in humans is anecdotal. A 1978 study of inhabitants of Okinawa¹ showed that energy intake among adults was roughly 80% of the Japanese average. Some researchers have held this up as evidence that restricting diet reduces age-related diseases and extends longevity. But gerontologist Sataro Goto of Toho University

in Chiba, Japan, points out that Japanese people have a daily energy intake nearly 20% less than the average of developed countries, and that the mean life-expectancy of Japanese women is 85 years — not significantly more than that of women in all developed countries.

Hungry for life

This hasn't stopped thousands of people all over the world intentionally going hungry in the hope of living longer and healthier. But no one yet knows if their punishing lifestyle will pay off. So an experiment designed to find out is being watched closely by researchers into ageing.

The CALERIE (Comprehensive Assessment of Long-term Effects of Reducing Intake of Energy) study, sponsored by the US National Institute on Aging (NIA), published the outcome of its first phase in April². Forty-eight men and women were randomly assigned to diets that would either maintain their weight, or reduce their calorie intake by up to 25% below the level required to maintain their weight. After six months, those on restricted diets showed lower body temperature and levels of insulin in their blood when they had not

had a meal, both characteristics of long-lived people and animals. This slowing of the metabolism explains why people on calorie-restricted diets don't starve to death: their body's energy requirements drop to meet the number of calories in their diet.

CALERIE is a departure in this hype-infested field, because the volunteers are of normal weight or slightly overweight, but not obese. Most studies so far have used obese individuals, muddying the waters with the health benefits of reducing obesity. The second phase of the study will begin this summer, observing 200 subjects over two years. But according to one of the present study's authors, physiologist Eric Ravussin of the Pennington Biomedical Research Center in Baton Rouge, Louisiana, there are already hints that calorie restriction affects ageing processes in cells.

Participants on restricted diets lost weight exponentially, and were still losing weight after six months. They had reduced insulin resistance and reduced levels of low-density lipoprotein cholesterol, high levels of which are risk factors for type 2 diabetes and heart

J. WESTRICH/ZEFA/CORBIS

disease, respectively. The researchers also measured significantly reduced damage to volunteers' DNA. "We asked if there is a metabolic adaptation over and above what is expected due to the weight loss," says Ravussin, "and the answer is yes, there is."

He says the findings are consistent with the theory that calorie restriction reduces metabolic rate and lowers the production of harmful molecules called reactive oxygen species, also known as free radicals. These strip electrons from other molecules in the body; this is thought to harm cells and contribute to ageing. This metabolic change, Ravussin believes, could be what causes calorie-restricted rodents to live up to 40% longer than their unrestricted counterparts.

Clearly it will be a long wait before the CALERIE study reveals a similar effect in humans, if it ever does. Meanwhile, the research community is divided over whether calorie restriction will extend lifespan in humans.

The most recent online edition of the journal *Biogerontology* asked a number of experts to address this question³, in response to a new theory of longevity championed by mathematical biologist Lloyd Demetrius of Harvard University. Demetrius argues that the main factor determining lifespan is not the rate of free-radical production, but the cells' ability to resist short-term fluctuations in critical metabolites caused by environmental stresses, so protecting those cells from damage.

Long haul

Demetrius believes this is determined by a creature's evolutionary history. Rodents are opportunistic species that experience periods of plenty punctuated by periods of scarcity. When there is lots of food they reproduce like crazy; when there isn't, they sit tight. They reach sexual maturity early, have a narrow reproductive window and large litters. Humans, and larger-bodied animals in general, mature late, have fewer offspring and a wider reproductive window. They are more able to move from resource-poor regions to resource-rich ones, and are more stable in the face of environmental perturbations⁴.

Demetrius's theory predicts that calorie restriction would extend lifespan in opportunistic species, giving them a chance to breed once conditions improve, but barely affect lifespan in more stable species. Some evidence supports this. As Linda Partridge of University College London's Centre for Research on Ageing notes: "Extension of lifespan by dietary restriction has been seen particularly clearly in animals with high reproductive rates — rodents, worms, flies." In contrast, experiments in rhesus monkeys have yet to show that calorie restriction increases lifespan⁵.

Believers in calorie restriction claim this is a methodological problem: monkeys live longer than rats, so the experiments take longer. Studies with rhesus monkeys begun in the United States in the late 1980s will not produce robust



Longevity island: Oto Yara, one of Okinawa's unusually large number of centenarians.

data for another 25 years, and human studies such as CALERIE will take even longer. In the meantime, they argue, the data from studies such as CALERIE support the idea that it could indeed work in people.

Such optimism means that the development of drugs aiming to mimic calorie restriction is well advanced, with some compounds already in clinical trials⁶. These drugs are designed to slow ageing and possibly extend life by triggering metabolic adaptation without the need for semi-starvation. Whether or not calorie restriction increases longevity, it does seem to ward off certain age-related diseases. For example, epidemiological studies have shown that low-calorie diets are associated with lower incidences of age-related ailments such as cancer and neurodegenerative diseases⁷.



Weighty issue: separating the benefits of calorie restriction from the dangers of obesity is tricky.

The field grew out of a change in thinking about ageing. Until about 15 years ago, ageing was thought to be an unregulated process resulting from the accumulation of cellular damage. Then genetic studies, particularly in the nematode worm *Caenorhabditis elegans*, showed that manipulating single genes could increase or decrease lifespan dramatically⁸. Later, lifespan-extending mutations were found in other organisms, including the fruit-fly *Drosophila* and rodents.

Many genes have been identified that are thought to play a part in determining lifespan, such as those involved in insulin signalling and stress resistance. One family of enzymes currently in the limelight is the sirtuins. The gene *SIR2*, after which the family is named, was one of the first longevity genes to be identified. Species from yeast to humans carry variants of it, and extra copies increase lifespan in yeast, worms and flies⁹.

Sirtuin something

Nobody knows how sirtuins might extend lifespan, but David Sinclair, who studies ageing at Harvard Medical School in Boston believes that they stabilize DNA and counteract the reduced fidelity of DNA-copying mechanisms in old cells. The Sir2 protein seems to work with a small molecule called NAD, which is involved in metabolism⁹, so the two together potentially explain the association between calorie restriction and ageing.

In 2003, Sinclair's team described 19 plant-derived molecules that activate sirtuins in yeast¹⁰. One of these, resveratrol, found in grape skins, has been reported to have anti-cancer and neuroprotective effects. The chemical's presence in wine has been touted as the explanation of the 'French paradox' — the low incidence of heart disease in southern Europe, despite a fatty diet. Sinclair and others believe it may also extend life.

Working with Rafael de Cabo, Donald Ingram and others at the NIA's Laboratory of Experimental Gerontology in Baltimore, Sinclair is studying the effects of resveratrol in non-obese mice. The study's results are not yet published, but de Cabo claims they look promising. There is already published evidence that resveratrol extends life in another vertebrate. In February this year, Alessandro Cellerino and his colleagues at the Scuola Normale Superiore in Pisa, Italy, revealed that the chemical not only extends the maximum lifespan of a short-lived fish called *Nothobranchius furzeri* by 60%, but also seems to protect the fish from neurodegeneration¹¹.

Cynthia Kenyon, a researcher in ageing at the University of California, San Francisco, is "very impressed" with the resveratrol results, but, she points out, there is confusion over what it actually does. "In animals, at least in some cases, it looks like the effects you see are dependent on Sir2," she says. On the other hand, she says, resveratrol has so far shown no effect on Sir2 function in human cells in culture. This makes

P. JONES GRIFFITHS/MAGNUM PHOTOS

D. HULSHIZER/AP



S. FRANKLIN/MAGNUM PHOTOS

Good vintage: a compound found in red wine called resveratrol might explain why the French have fatty diets but long lives.

it unlikely that the chemical acts directly on the Sir2 protein, although it might target other biochemical processes that interact with Sir2.

How it works may not matter in the end, she says. But in the absence of any conclusive evidence that sirtuins extend life in mammals, or that resveratrol activates sirtuins in human cells, Sirtris Pharmaceuticals, a company co-founded by Sinclair and based in Cambridge, Massachusetts, has screened half a million compounds and found some that activate human sirtuins in the test tube.

Natural experiments

According to Sinclair, Sirtris has unpublished animal data showing that some of these sirtuin activators affect markers of longevity in the same way that calorie restriction does in rodents and in the humans in the CALERIE study — for example, by reducing insulin levels. One of the molecules is already in clinical trials, although Sinclair warns it will be several years before Sirtris can show that this is an effective way to treat age-related disease, let alone extend life.

Kenyon is co-founder of another Cambridge-based company pursuing anti-ageing drugs, Elixir Pharmaceuticals. Elixir is interested in sirtuins, but is taking a broader approach. Kenyon's group discovered that activity in biochemical processes controlled by

insulin and another hormone called insulin-like growth factor 1 controls ageing. That work has led them to test molecules that affect those mechanisms, which control the body's use of glucose. The team is also investigating drugs that block the effects of ghrelin, another hormone that controls the body's glucose balance.

"We are thinking about getting drugs approved not for ageing, but for beneficial effects on age-related diseases."
— Cynthia Kenyon

"The way we are thinking about getting drugs approved is not of course for ageing, but instead for beneficial effects on age-related disease," says Kenyon. She envisages that people will start taking them once they have been approved for the latter — probably within the next five to ten years — and a huge natural experiment will be launched. "If they really are drugs against ageing, two things should happen," she says. "One, they should have efficacy in additional disease settings, and two, after a while you might start noticing that people taking the drugs seem younger."

Just such an experiment is underway with the drug metformin, which enhances insulin action and was approved for treating type 2 diabetes in 1995. Since then, studies in rodents have shown that metformin can reduce tumour load, and reduce fasting insulin level and body temperature without significantly affecting body weight or food intake. But some researchers are doubtful about its usefulness as an anti-ageing drug, because it has side effects

such as a small risk of a build-up of lactic acid in the body which, rarely, can be fatal.

We may never be sure whether calorie restriction extends human lifespan, says Ingram. He argues that it is unlikely that a test of the effects of lifelong restriction in humans under controlled conditions will ever be done, for reasons of cost and practicality.

In the meantime, some believe that researchers into ageing must follow any lead, because of the social and economic issues posed by a rapidly ageing population. A group of experts led by demographer Jay Olshansky of the University of Illinois, Chicago, is calling on the US government to increase its funding of basic ageing research, with the initial goal of delaying all age-related diseases by about seven years — a goal they regard as realizable for generations living today.

Laura Spinney is a science writer living in London and Paris.

1. Kagawa, Y. *Prev. Med.* **7**, 205–217 (1978).
2. Heilbronn, L. K. et al. *J. Am. Med. Assoc.* **295**, 1539–1548 (2006).
3. Le Bourg, E. & Rattan, S. I. S. *Biogerontology* doi: 10.1007/s10522-006-9010-5 (2006).
4. Le Bourg, E. *Biogerontology* doi: 10.1007/s10522-006-9014-1 (2006).
5. Kayo, T., Allison, D. B., Weindruch, R. & Prolla, T. A. *Proc. Natl Acad. Sci. USA* **98**, 5093–5098 (2001).
6. Ingram, D. K. et al. *Ageing Cell* **5**, 97–108 (2006).
7. Roberts, C. K. & Barnard, R. J. *J. Appl. Physiol.* **98**, 3–30 (2005).
8. Kenyon, C. et al. *Nature* **366**, 461–464 (1993).
9. Sauve, A. A., Wolberger, C., Schramm, V. L. & Boeke, J. D. *Annu. Rev. Biochem.* **75**, 435–465 (2006).
10. Howitz, K. T. et al. *Nature* **425**, 191–196 (2003).
11. Valenzano, D. R. et al. *Curr. Biol.* **16**, 296–300 (2006).

BUSINESS

Burst of energy

More and more venture capitalists are backing clean technology in the United States, but will it take off? Virginia Gewin reports.

For years, venture capitalists spurned Lee Lynd's plans to make lucrative ethanol out of agricultural waste. But in March, that suddenly changed. Lynd, an engineer at Dartmouth College in New Hampshire, got US\$4 million from venture capitalist Vinod Khosla to expand his company, Mascoma, into one of the leading US ethanol producers.

Lynd's idea has been buoyed by the tide of money flowing into clean energy as US investors wise up to its potential, in the light of high oil prices and rising public concern over energy security.

From venture capital to regular investments by staid financial institutions and pension funds, money is now pouring into technologies such as solar, wind and biofuel. The amount of venture capital has risen from very low levels a decade ago to almost a billion dollars last year (see graph). That's just over 4% of all such capital in the United States, according to Clean Edge, a market-analysis firm based in San Francisco. Last year, some 160 new US companies started in the sector. And three of the largest initial public offerings of technology stocks in 2005 were for solar energy companies: Q-Cells, SunPower and Suntech Power raised some \$800 million in total.

Boom time?

But the upsurge in interest from the money men is not without its pitfalls. Some analysts are already asking whether the splurge has the makings of a bubble. They wonder if the United States, where public demand for alternative energy sources lags behind that in Europe and elsewhere, can support so many new companies. And with little public investment in energy-related scientific research over the past 20 years, they want to know where the fresh ideas are coming from that will support successful ventures.

So far, many of the clean-energy companies are being built on technology derived from other high-tech sectors, such as electronics or biotechnology. The photovoltaic cells made by solar-power companies, for example, use silicon wafers developed over many years by the electronic industry. And biotechnology has produced the enzymes that can efficiently convert cellulose material from plants into ethanol.

After years of neglect, the Bush administration is now trying to push this technology along by doing more energy-related research in government laboratories. President George W. Bush's Advanced Energy Initiative, which was announced earlier this year, would increase the Department of Energy's budget for this kind of work by 22%, to \$771 million, next year.

But Bill Green of Vantage Point Venture Partners, a Silicon Valley investment firm that backs small companies in the sector, says a lack of research dollars is not what is holding them back. He says the main problem is the ability to scale ideas up and launch them into the energy market.

Indeed, companies such as Mascoma are mainly focused on increasing the efficiency of existing production abilities. The cost of enzymes for biomass-based ethanol production has decreased 30-fold since 2001, but high processing prices remain the biggest challenge in taking bioethanol to market. Mascoma has addressed this by developing an efficient pretreatment process that reduces waste and produces saleable by-products.

The government is aware of the problems small energy companies face in scaling up. Floyd Kvamme, co-chairman of the President's Council of Advisors on Science and Technology, is currently drawing up a set of policy recommendations for nurturing such companies. He says the government should do more to promote demand for renewable energy.

Market watchers agree that it is demand for renewable energy, more than the supply of new ideas, that will determine the fate of the young companies. Michael Liebreich, chief executive of New Energy Finance based in London, says he doubts that vastly innovative technologies

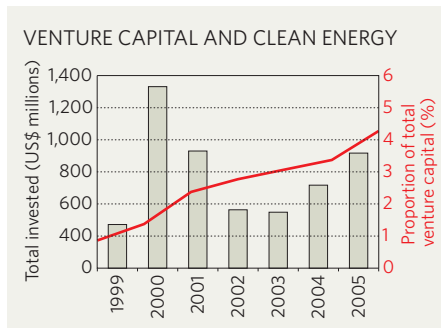


will be needed. He says there are some areas, such as advanced materials for solar panels, enzymes for biofuel processing, software for grid management and battery development, in which new ideas could yield the high returns sought by venture capitalists. But he thinks they should tread carefully. "You don't want a situation where too many venture capitalists are chasing too few good deals," he says, noting that this could produce a boom-bust cycle, in which too much hype encourages low-quality investments that yield disappointing returns.

Dumb money

One reason for caution is the shortage of entrepreneurs or investors who understand the nuances and complexities of the sector. "People want to see a lot happen," says Lynd, "but there are not that many competent people to make it happen." And some analysts think that the venture capitalists are already overfunding projects. "There is a lot of dumb money out there," contends Peter Fusaro, founder of Global Change Associates, an energy investment consultancy in New York.

The National Renewable Energy Laboratory (NREL) is the principal research institute for developing alternative-energy technology in the United States. Many advances in photovoltaic cells, for example, have sprung from the laboratory at Golden, Colorado. Although the lab's primary focus is on research, it has also actively sought to connect investors with technology. But its efforts are hampered, lab officials say, by government red tape. "If you are on a research and development contract with NREL, you can't spend money on market





D. A. FRASER

These bioreactors produce ethanol from cellulose, but will they be built on an industrial scale?

analysis or validation,” says Lawrence Murphy, the lab’s manager of enterprise development programmes.

Critics say that such rules are an immense impediment to technology transfer. Dan Reicher was assistant energy secretary in the Clinton administration and now runs New Energy Capital, a firm in New England that backs renewable energy projects. He says that there is far too little coordination between government and the private sector.

“The Department of Energy has got to shift from research to deployment issues,” says Mark Sinclair, deputy director of the Clean Energy States Alliance, a non-profit collaboration of 17 state funds to promote markets for clean energy, headquartered in Vermont. Sinclair would like to see federal action that would encourage, for example, the adoption of photovoltaics by home builders.

“We are in policy wasteland,” agrees Green. He points out that even though 22 states require electricity firms to generate a minimum percentage of power from renewable sources, they each have different rules and targets, which makes life hard for companies.

Others are more sanguine, arguing that good ideas will work their way through to the market in the end. Despite the lack of a national energy policy, business will see plenty of opportunities for high returns as long as the price of oil and gas remains high. “The capital markets are driving this train,” says Fusaro. ■

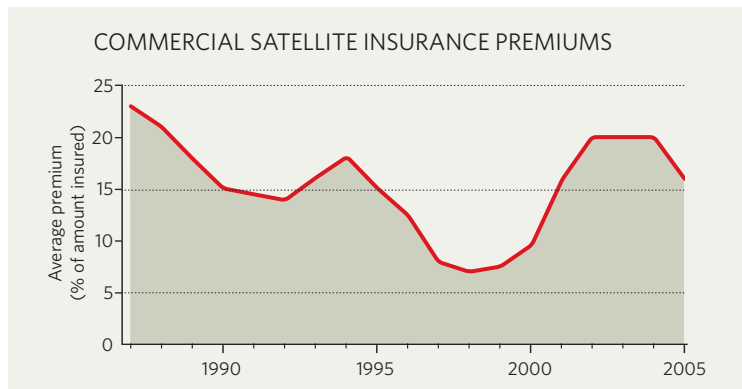
IN BRIEF

INDIAN OVERDRIVE IBM has announced a huge push of people and resources into India, saying that it will invest US\$6 billion there over the next three years. Its plans were unveiled by IBM chairman Sam Palmisano and Indian president Avul Kalam to a crowd of 11,000 cheering employees in Bangalore on 6 June. They include the expansion of existing computer laboratories in that city, as well as the establishment of a global telecommunications laboratory in New Delhi. The computer firm, headquartered in Armonk, New York, already employs some 43,000 people in India — more than it does in any other country outside the United States.

BACK ON SALE The US Food and Drug Administration (FDA) says it will allow the multiple sclerosis drug Tysabri to return to the market — under strictly monitored conditions. Tysabri (natalizumab) was approved in 2004 but pulled from sale in February last year after its manufacturers, Biogen Idec of Massachusetts and Irish biotechnology company Elan, said that two patients had died in clinical trials of the drug. Shares in the two companies dipped on the 5 June decision, however, owing to the harshness of the proposed restrictions: Biogen Idec dropped \$2.32 to \$45.39, and Elan fell \$2.46 to \$16.52.

GENERIC GO-AHEAD The FDA, under pressure from the courts (see *Nature* 441, 23; 2006), has approved a copycat version of a biologic — a drug made from living cells — for the first time. Omnitrope, which won European approval in April, is a generic version of Pfizer’s Genotropin (a human growth hormone) and has been approved for the treatment of growth disorders. Sandoz, the German Novartis unit that makes Omnitrope, labelled the decision a “breakthrough” for makers of generics. But the FDA says its decision does not set a precedent for approval of generic biotech drugs, because most of them are regulated by a different statute from that applying to growth hormones.

MARKET WATCH



If you’re an insurance company that finds house fires and fender-benders a little bit mundane, there’s always the space business.

With premiums of, say, US\$40 million a shot and payouts that can be ten times that, insuring commercial satellite launches is not for the faint-hearted.

Yet after a series of accidents that took these premiums to highs of as much as one-fifth of entire project costs, premiums slipped back last year (see graph).

“Insurers are in business to make a profit, so if you have a period when the business is profitable, others will come into the market,” explains Rick Hauck, former chief executive of AXA Space, a satellite insurer based in Bethesda, Maryland.

Premiums rose sharply after the solar panels failed on six Boeing 702 communications satellites in 2000 and 2001 — costing insurers some \$1.5 billion. High premiums subsequently put a damper on commercial satellite launches, which were already suffering from the collapsing fortunes of the telecommunications industry.

But an assessment by the US Federal Aviation Administration, published last month, suggests that premiums are finally starting to dip again.

No one knows whether the fall will be sustained. “It is in the hands of events,” says Hauck. “If there are profits, others come in to compete and that puts pressure on the rates.” ■

Colin Macilwain

Women editors: change comes from focused action

SIR — Your News story “Societies spurn women editors” (*Nature* **440**, 974–975; 2006) raises valid concerns about the lack of women in leadership positions in many scientific societies and on many editorial boards. The Society for Glycobiology, of which I am currently president, has a membership of chemists, biochemists and biologists. The majority of our members are male, but the society takes a direct route to ensure participation of women in all our activities.

Our leaders and members, as well as the editorial board of our journal, *Glycobiology*, actively encourage the election of women to leadership positions. Women currently occupy three of the six positions on the society’s board of directors. Three of the past ten society presidents have been women, and women make up about 20% of the editorial board. The programme committee has ensured participation of women as platform speakers and session chairs at our annual national meeting. In addition, our policy is to include short talks by two or three junior scientists in every session of the meeting, to allow exposure of their work in a national forum — a policy that helps junior scientists of both genders.

I cite these figures to highlight how a concerted effort to be inclusive can yield meaningful change in leadership positions. Focused, explicit and determined actions to change the balance of participation by women in scientific societies may be required to effect rapid change, rather than the hope that change will eventually occur over time.

Linda G. Baum

Department of Pathology, UCLA School of Medicine, 10833 LeConte Avenue, Los Angeles, California 90095, USA

Women editors: we need more female scientists

SIR — Your News story “Societies spurn women editors” (*Nature* **440**, 974–975; 2006) suggests that there is ongoing and systematic bias against women for the position of editor of *Evolution*. As vice-president of the Society for the Study of Evolution and associate editor of *Evolution*, I strongly believe this is not the case. Several women were identified as potential editors-in-chief of *Evolution*: women with strong research records, broad visions for the field and wide support. But not one of those approached was interested in the position.

Why are women declining to be considered? In the face of the many obstacles to women in science, those of us who have established successful research careers have done so by

learning to say “no”. The task of editor-in-chief is onerous, time-consuming and largely thankless. Potential candidates recognize that saying “yes” to this particular job comes at an enormous personal cost — to one’s research career, students and family.

It is important to recognize that women in science do a disproportionate share of service because of the desire for gender equality on committees. Expecting us to do more is not necessarily in the best interests of women or science. The paucity of female editors points to the need to encourage the scientific careers of women: only when there are more senior women scientists are we likely to see this problem go away.

Sarah P. Otto

Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada

Women editors: nominees turned down *Evolution* job

SIR — As president of the Society for the Study of Evolution, I object to the claim that women were not adequately considered for the position of editor-in-chief of the journal *Evolution* (“Societies spurn women editors” *Nature* **440**, 974–975; 2006). The society is not sexist, nor do we reject women for senior positions. Of the past six presidents, three have been women. Our governing body has eight men and seven women, including our current chief executive and financial officer. Some 42 (29%) of our associate editors since 1995 have been women, compared with 23% of our membership.

It is true that we have had only one female editor-in-chief. However, five women were approached about their interest in the position before Mark Rausher was appointed, and all declined. The selection process involved a committee of six instead of three as stipulated in our by-laws, but this widened input and involved more women — including Theresa Markow. She raised no procedural issues until after the new editor-in-chief had been selected, agreed to serve and started to appoint other editors. The officers and council responded promptly to these issues in early 2006, but voted to approve the new editors rather than repeat the search.

Ironically, Markow’s resignation deprives evolutionists of a woman at the most senior position in our society.

I must also ask: why has *Nature* never had a female editor-in-chief?

Don Waller

Department of Botany, 430 Lincoln Drive, University of Wisconsin, Madison, Wisconsin 53706, USA

We stand by our description of the selection procedure — Editor, *Nature*.

‘Referee factor’ would reward a vital contribution

SIR — At a time when academia is increasingly focused on measures of performance, there is no provision for those who referee articles, even though the process is onerous, time-consuming and critical to the advance of science. For the system to be fair, all scientists should be refereeing two to three times as many articles as they submit.

The solution is simple. Some journals already acknowledge referees at the end of each year, but if this process were standardized for all journals, each stating the number of manuscripts seen by individuals, the data could be compiled by a suitable mechanism such as Google Scholar. The results could ultimately be translated into a ‘referee factor’, consisting of the sum of the impact factors for the respective journals multiplied by the number of articles reviewed. Multiple versions of the same manuscript probably need not be factored in, as the work involved in assessing revisions is considerably less than that for previously unseen contributions. The ‘referee factor’ could be built into standard assessments of performance, acting as an incentive for people to review manuscripts and otherwise making attitudes apparent.

Rory Wilson

Biological Sciences, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, Wales, UK

See Nature’s web debate on peer review at www.nature.com/nature/peerreview/debate/index.html

Still light-years away from articulating the infinite

SIR — The Universe may perhaps be infinite, but even to distinguish this from a very large finite Universe would seem to be scientifically impossible. Yet objects such as comet Encke, used to illustrate Peter Coles’ book review (“From here to eternity” *Nature* **441**, 285; 2006), hardly encourage us to consider an infinite universe to be reasonable. Encke is a body in the Solar System, a few light-hours away. In fact, it is one of the closest comets to the Sun.

You might have pictured a distant galactic cluster, such as the Great Wall — perhaps 200–300 million light years away. But even that would be a finite distance, no closer to infinity than a walk across the road.

Ian Stewart

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

Contributions to Correspondence may be submitted to corres@nature.com. They should be signed by no more than three authors; preferably by one.

BOOKS & ARTS

In your own image

Care must be taken when looking for natural selection to explain the evolution of human behaviour.

Before the Dawn: Recovering the Lost History of Our Ancestors

by Nicholas Wade

Penguin: 2006. 320 pp. \$24.95

Kenneth M. Weiss and Anne V. Buchanan

In *Before the Dawn*, journalist Nicholas Wade explores the “lost history of our ancestors” from a genetic viewpoint. He presents a compilation of scenarios that are meant to explain the evolutionary origins of human behaviour and social structure during the past 50,000 years, discussing archaeological findings but mainly focusing on how genes can reveal the ‘hidden’ past that can’t be inferred from fossils or archaeology. He uses examples ranging from the first wearing of clothes and the origin of hairdressing to the evolution of language, race and intelligence.

These inherently interesting and smoothly presented tales of progress towards the human present will doubtless captivate many readers. However, what could have been a tempered and timely treatment of an important subject is, in our view, regularly undermined by Wade’s determination to find simplistic natural selection behind every trait, and by a lack of attention to issues that are known to inhibit a credible understanding of complex traits, never mind their evolution.

Wade’s explanations commit various well-known errors, such as equating correlation with causation and extrapolating from individual traits to group characteristics. Often his arguments and trait choices are laden with Western-oriented value judgements. The following are a few examples of the kinds of problematic scenario that can be found in the book. Wade suggests that until recently racial variation has not been the subject of scholarly pursuit, and confidently asserts that there are five “major” races, identified statistically by neutral markers (from a genome that he assures us is 97% “filler”), yet “the alleles involved in differentiating the human population are likely to be of the selected kind not the neutral kind”. He also states that stressing genetic racial differences as he does is “objective” and scientific, but stressing human similarities is “political”. Wade argues that Europeans resist ‘mad cow disease’ because their ancestors were selected for cannibalism. He also says that Jews were selected for higher intelligence than other peoples because of



P. PARKS/AFP/GETTY

Born to win: does the ability of some Chinese people mean there is a gene for table-tennis?

the calculational demands of money-lending. He suggests that high intellectual skills are a genetic adaptation that occurred only after the origin of settled societies in places such as Europe. And he says that the Chinese as a “race or ethnic group” excel at ping-pong, which should encourage researchers to look for a genetic explanation.

Extraordinary claims demand extraordinary evidence. In *The New York Times* on 15 January 2006, Wade warned against journalists being too ready to accept “overstated or wrong” claims from the science literature, but in too many places where it makes a difference he has ignored his own advice. A journalist doesn’t create facts, but he does select what to repeat and how to colour it, and Wade is long on speculating about what “is reasonable to assume”, and short on circumspection of his own, or anthropologists’, yarn-spinning. Most of the scenarios he reports have not been rigorously tested, nor is it clear how they could be. The book has many internal inconsistencies, and one can easily find contrary evidence or readily construct alternative ‘just so’ stories that invoke the same genetic scenario and the same kind of reasoning.

How could this subject be better treated, without denying the importance of genes in human traits? For a start, evolutionary arguments

should be based on sufficiently credible, consistent and compelling scientific evidence. It is easy to claim that a trait is due to natural selection, but responsible selection-based arguments should have substantial experimental mechanistic support, at least for the fact of selection. That’s not the state of most current evidence. Indeed, after 50 years of investigation, we can’t convincingly demonstrate selection for most of the red-blood-cell diseases, other than sickle-cell anaemia, that are probably coevolving with the strong selective force of malaria. Other best-case scenarios for human genetic adaptation, such as adult lactase persistence and skin colour, are also incomplete. Explaining selection is particularly problematic for behavioural traits because of the powerful role of culture and facultative ability, which is probably what human evolution really favoured. Human phenotypic changes can far outpace genetic ones, making it challenging to know whether such traits are even genetic, much less what they ‘evolved for’ millennia ago.

In addition, assertions of genetic causation should be built on what is already known about the difficulties of explaining complex traits, including behaviour or intelligence. The extensive literature documenting the subtleties of such traits undermines simplistic ‘evolved for’ scenarios, but Wade largely ignores it. The

aetiology of complex traits is influenced by environmental factors as well as variation at multiple genes, greatly attenuating the causal impact of individual genes. We are far from understanding either the genetic architecture or the evolution of complex biological traits, even in the best data from experimental organisms unaffected by the blur of culture. Intensive gene mapping has typically failed to identify more than a fraction of even the genetic variation, much less all the variation, in such traits. The effects of experimental genetic manipulation in laboratory animals routinely vary significantly even among the few strains tested, and the life experiences of litter-mates, twins, inbred animals and clones are far from identical. Despite this sobering knowledge, Wade claims example after example of 'genes for' traits.

But why not just enjoy the sport of fanciful speculation, even if the arguments leak like sieves? Because it's not just sport. Positions on genetic determinism often correlate with

social politics, and few of us are neutral or even changeable on the issues. Wade recognizes that his ideas may not be acceptable to everyone but warns that "to falter in scientific inquiry would be a retreat into darkness". He seems to be warning, appropriately enough, against benighted political correctness. But we should never become casual about how comparable 'slopular' science and very similar speculative evolutionary reasoning by leading scientists fed a venomous kind of darkness not too many decades ago. Wade's post-hoc tales often put him in step with a long march of social darwinists who, with comfortable detachment from the (currently) dominant culture, insist that we look starkly at life in the raw and not blink at what we see. But given today's limited understanding of complex traits, too often what one sees is oneself. ■

Kenneth M. Weiss and Anne V. Buchanan are in the Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

intend to make it required reading for my astronomy students (and strongly recommend the same for similar courses), mostly because it provides the right approach to the phenomenology of the Universe. Apart from the elusive gravitational waves, astronomers learn most about our Universe by gathering and analysing electromagnetic radiation from it. The art of astronomy is to extract the maximum information carried to us from the stars by travelling photons: arrival direction, arrival time, energy and state of polarization. And, of course, to gather as many photons as possible.

We've done this for centuries with bigger and bigger telescopes on the ground, but "we live at the bottom of an ocean of air", as Galileo's favourite pupil, Evangelista Torricelli, remarked with astonishing insight. The bulk of the spectrum can only be studied properly from space. For the first time in the history of humankind, astronomy is possible out there, and we've had close encounters with planets and comets in the Solar System. Europe has sent probes to planets and comets (Huygens, Mars Express, Rosetta, Venus Express and so on) through ESA's science programme, and these are discussed in Woltjer's book.

Woltjer has also been a leading figure in space-astronomy planning for Europe, working with Roger Bonnet in crafting the first two long-range ESA science plans. It is a pleasure and an honour to report that this has continued in ESA's 2004 Cosmic Vision plan, based on the science proposed by a community that has trebled in size over the past 20 years.

Europe's Quest for the Universe tells the story of European astronomy with passion, immediacy and candour, and with as much mastery of science as it has personality. ■

Giovanni Bignami is at the Centre d'Etude Spatiale des Rayonnements, 9 avenue du Colonel Roche, 31028 Toulouse, France.

Reaching for the stars

Europe's Quest for the Universe

by Lodewijk Woltjer

EDP Sciences: 2006. 328 pp. €35

Giovanni Bignami

"The progress of science depends on the technological development of its instrumentation," claims Lodewijk Woltjer in the opening line of *Europe's Quest for the Universe*. From Galileo to Edwin Hubble and Riccardo Giacconi, this has been especially true for astronomy. But, as Europe's former research commissioner Philippe Busquin adds in his preface: "The star-studded sky acts as a source of wonderment and inspiration for our thoughts and dreams." And that's lucky, in these days of obsessive, applications-oriented political pressures on science.

Both from the ground and from space, astronomy has been a success story for Europe, and one that needed to be told. Descriptions of bold endeavours can come from leaders in the field and be accurate yet not impersonal, or they can come from observers without a direct influence on the outcome. Woltjer was director-general of the European Southern Observatory (ESO) for eight years and has been a keen observer of astronomy in Europe and across the world. He writes, then, both as a leader and a commentator. This turns out to be a perfect recipe for the history of the varied and variable discipline of space science, in a continent that has more nations than big telescopes, and where the number of active space astronomy missions is close to the number of member states of the European Space Agency (ESA).

From the ruins of the Second World War, Europe, like Japan, has emerged to become a

world-class power in the peaceful and cultural endeavour to understand our Universe. It all came together at the same time, in those magic years at the end of the 1950s when everything must have seemed possible: we had CERN, the European particle-physics laboratory; the European Space Research Organisation (ESRO); and, of course, the ESO. I wonder whether Europeans today would be capable of pulling off a similar coup in our extremely affluent society?

But a merely historical reading would be a limited and limiting one for Woltjer's book. I



The European Southern Observatory operates the Very Large Telescope array at Paranal in Chile.

A growing urban problem

Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals

by Michael Batty

MIT Press: 2005. 648 pp. \$60, £38.95

Frank Schweitzer

Urban growth is one of the biggest challenges for humankind in the twenty-first century. The world's urban population is at present estimated to be growing by about 50 million people per year. This growth is still almost exponential, and saturation may be reached only by the middle of the century. So there is much discussion about estimates of land consumption, the manageable size of megacities, and simple ways to structure, if not to control, this vast spread of urbanization.

For this, a better understanding of the spatial dynamics of urban growth should be a precondition. In fact, urban planning, urban geography, urban economics and other related disciplines have put quite some effort into this problem. Although they provide insight into some details, they do not present the big picture of generic mechanisms underlying these complex dynamics. So it is reasonable to consider whether the theory of complex systems developed in different scientific areas over the past 30 years may provide a suitable toolbox for gaining such an insight.

In *Cities and Complexity*, Mike Batty, an expert in the area of urban modelling, unfolds the different aspects of urban change in relation to complex systems theory. The focus is on models and computer simulations: there is no introduction explaining what we already know about cities and growth, or any attempt at orientation on the phenomenological side. This is good and bad at the same time. Some readers will certainly miss a clear relation between cities as we know them and the rather abstract models and results for specific dynamic features, such as segregation or polynucleated growth. Others, however, will appreciate the insights into generic features of urban dynamics that can only be highlighted at the abstract level used. As the author puts it, the focus is "largely on experiments with models that provide us with analogies as to how cities develop and evolve".

These analogies are investigated on very different modelling levels relating to the different scales of urban dynamics. Models of fractal growth show analogies to the morphological structure of urban aggregates, and models of moving agents resemble the motion of pedestrians on streets and inside buildings. Many of these models use cellular automata, which allow a broad range of topics such as land-use patterns or regeneration of urban areas to be addressed. A 'desktop simulator' is described in more detail in one of the chapters, tempting



S. GALLUP/GETTY

The expanding sprawl of cities such as Athens highlights a need to understand urban growth.

the reader to play with the ideas discussed in the book. Unfortunately, the plan to put examples from the book together with related material on a website (www.complexcity.info) is yet to be realized.

The book certainly succeeds in making the topic of urban growth accessible and interesting to those already familiar with formal modelling of complex systems. However, practitioners in urban planning may find some of the concepts and models too abstract and 'academic' to give solutions. But at least they

should agree that this complex-systems perspective sheds new light on to the dynamics of urban evolution by highlighting relations to seemingly distant phenomena, such as swarming or epidemics. Apart from the necessary technical details of the models, the author gives a number of illustrative examples and facts to elucidate his viewpoint, widening the appeal of the book to a broader audience. ■

Frank Schweitzer is chair of systems design at ETH Zürich, Kreuzplatz 5, 8032 Zürich, Switzerland.

A healthy interest

Doctor Franklin's Medicine

by Stanley Finger

University of Pennsylvania Press: 2006.
400 pp. \$39.95, £26

W. F. Bynum

If the creation of the American Republic can be taken as a high point of the Enlightenment, then Benjamin Franklin (1706–1790) is a central figure. Unlike many of the other Founding Fathers, he came from a humble background and had a gentle sense of humour. Self-taught, he made his initial reputation as a printer and publisher, especially of *Poor Richard's Almanac*, a phenomenally successful annual in a crowded market. In the almanac, Franklin's alter ego, Poor Richard, offered advice and homespun wisdom about health and much else besides. In his own persona, Franklin was

to do the same for the rest of his long life.

In *Doctor Franklin's Medicine*, Stanley Finger has assembled Franklin's many interactions with medicine, health and doctors. Given the attention that the iconic Franklin has already attracted from historians, there are few surprises here, but the impact of the whole package is formidable. Most doctors would have been well satisfied to have so changed the face of medicine; Franklin did it almost incidentally, in the midst of a busy life as a natural philosopher, diplomat and man of the world.

Franklin lived most of the second half of his life in England and France, on official business first on behalf of Pennsylvania, and then of the new nation. He was known abroad primarily as an electrician, a word that in those days carried rather different connotations. Through his famous experiment with a kite and lightning,

he did for electricity what Newton had done with gravity: related the heavens to the Earth. It was as an electrician that Franklin was lured into the practice of medicine: he was several times asked to electrify patients with nervous disorders. Always suspicious of grand speculation, and a careful observer, he never claimed more for the therapeutic potential of his brain-child than his own experience warranted.

Several of his interactions with medicine have biographical poignancy. He advocated inoculation for smallpox (he died shortly before Edward Jenner introduced vaccination), but lost one of his own uninoculated children to the disease. Ever alert to the dangers of lead poisoning (among printers who used lead type, as well as other occupational groups), Franklin collaborated with George Baker, the British physician who exposed its high levels in cider and other alcoholic drinks kept in lead vats. Franklin himself loved Madeira and port, two

likely sources of toxic levels of the metal. He suffered from gout and bladder stones, a probable consequence of the poison to which he helped alert the public.

There were other inconsistencies in his philosophy of health. He advocated the health-giving properties of fresh air, although as a social man he thrived in the clubs of urban centres such as Philadelphia, London and Paris. He also preached the virtues of exercise, believing that swimming was an undervalued activity, and swam even after he became very corpulent in his old age.

Other medical contributions were unambiguous. Franklin invented and wore bifocals, and used an ingenious mechanical arm to grasp books and other objects on high shelves, a natural consequence of his large private library. He was a founder of the first public hospital in America, the Pennsylvania Hospital. So close was his relation to medicine that

the French occasionally assumed that he was medically qualified; in fact, 'Doctor Franklin' had several honorary degrees but no formal medical training. Nor did he apparently need it. He moved easily in medical and scientific circles, respected by the French and welcomed in Britain even after the United States had severed its ties with the mother country. Franklin had signed the Declaration of Independence, but he was always a man of peace. As he wrote in 1783: "There never was a good war, or a bad peace."

W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine, University College London, London NW1 2BE, UK.

MORE ON BENJAMIN FRANKLIN

The First Scientific American: Benjamin Franklin and the Pursuit of Genius

by Joyce E. Chaplin

Basic Books: 2006. 362 pp. \$27.50

Home from home

History is brought to life at Benjamin Franklin's house in London.

Colin Martin

Arguably the most famous American in the Age of the Enlightenment, Benjamin Franklin achieved scientific fame by flying a silk kite with a wire rod at one end and a key at the other into a thunderstorm in Philadelphia in 1752. He thus demonstrated that lightning consists of flashes of electricity.

During his long life, Franklin made many contributions to several branches of science. He found research a welcome respite from his role negotiating the turbulent complexities of contemporary diplomacy, including the repeal of stamp duty, which the British parliament had imposed on its American colonies in 1765.

This year marks the 300th anniversary of Franklin's birth, the cue for activities around the world (see www.benfranklin300.org). It has been celebrated in London by the public opening of the 1730s Georgian townhouse, at 36 Craven Street, where he lived for 16 years as an agent for the Pennsylvania Assembly. The newly restored house serves as a museum and education centre (see www.benjaminfranklinhouse.org).

During his years in London, Franklin invented bifocal spectacles; developed a fuel-efficient fireplace draught called the Franklin stove; installed a new, more effective lightning rod on the dome of St Paul's Cathedral; and invented the glass armonica (or harmonica), with its haunting

sound described as "the voice of angels". He worked with Joseph Priestley on experiments that led to the discovery of oxygen; recorded the effects of the Gulf Stream and other ocean currents; investigated canal depths and their implication for transport; and demonstrated the veracity of the old adage that oil calms

account of coagulation. Hewson lived at the house from 1770 to 1774 after he married Polly Stevenson, the daughter of Franklin's landlady. The medical-history room in the science study centre continues this tradition with ingenious models and touch-screen computers to encourage schoolchildren to think about how the human body works.

The house is interpreted for visitors in an evocative 'museum as theatre' tour, which incorporates state-of-the-art audiovisual display techniques and an actress in eighteenth-century dress who plays the role of Polly (see picture). Polly's interest in Franklin's diplomacy and science made her one of his closest confidantes. "After writing six folio pages of scientific philosophy to a young girl, is it necessary to finish such a letter with a compliment?" asks Franklin during an audiovisual presentation. "Is not such a letter of itself a compliment?"

Despite his skilled diplomacy, and his satirical *Rules By Which a Great Empire May be Reduced to a Small One*, written in 1773 to warn the intransigent British of the dangers of taking a hard line against hot-headed American colonists, Franklin failed to avert the War of Independence. In 1775, he hurriedly left Craven Street. By the time he had returned to Philadelphia, the American Revolution had begun. He died in 1790 at Passy near Paris, with Polly and Franklin's own daughter Sally at his bedside. Colin Martin is a writer based in London.



Tour de force? Polly acts as guide in Franklin's London home.

troubled water, on a windswept pond in the London suburb of Clapham. Some of these achievements are recreated as hands-on experiments for school children at the centre as an introduction to scientific methodology.

During the recent restoration of the house, some 1,200 human bones were found in the basement. They were identified as remains from an anatomy school founded at the house by William Hewson, a British anatomist who identified the role of fibrinogen in 1770 and gave the first valid

STRUCTURAL BIOLOGY

Images from the surface of HIV

Dennis R. Burton

Human and monkey immunodeficiency viruses are studded with 'spikes' that enable them to infect cells. Structural studies reveal that these spikes are tripod-like assemblies that cluster on the virus surface.

Newly produced HIV particles are in limbo between life and death, their fate determined by whether protruding structures — 'spikes' — on their surface make contact with receptors on the surface of white blood cells, in particular a subset known as CD4⁺ T cells. If a productive contact is made, a molecular sequence is triggered that results in the genetic material of the virus being injected into the target cell. Soon after, the viral genome begins to dictate events in the cell, leading to the production of many more virus particles, and full-blown infection gets under way. But if there is no productive contact between the viral spikes and target cells, then the spikes seem to lose their function and the HIV particles effectively die. The viral spike is thus central to HIV infection, making it a prime target for potential HIV vaccines, with the idea that blocking the spike–receptor contact using antibodies will prevent infection. On page 847 of this issue, Zhu *et al.*¹ report a low-resolution (at about 3 nm) structure of the spikes and their distribution on the virus surface. There are some real surprises.

The HIV genome, and that of its monkey counterpart, simian immunodeficiency virus (SIV), is enveloped in a lipid membrane derived from an infected cell. Virally encoded spikes, often known as envelope spikes, are anchored in this membrane and protrude from the surface of the particle. They are composed of two interacting proteins: gp41, which spans the membrane, and gp120, which sits on gp41 at the surface of the virus particle. Both proteins are glycosylated (that is, have sugar groups attached); indeed, gp120 is one of the most heavily glycosylated proteins known, with about 50% of its molecular mass being carbohydrate. This sugar coating helps HIV to elude the immune system because the carbohydrate is derived from the infected cell and is poorly recognized by antibodies. But HIV also has other means of distracting the immune system, including loops of highly variable sequence that extend from the surface of gp120. These loops present an ever-changing target for antibodies and create a huge headache for vaccine designers.

Information about the envelope spike struc-

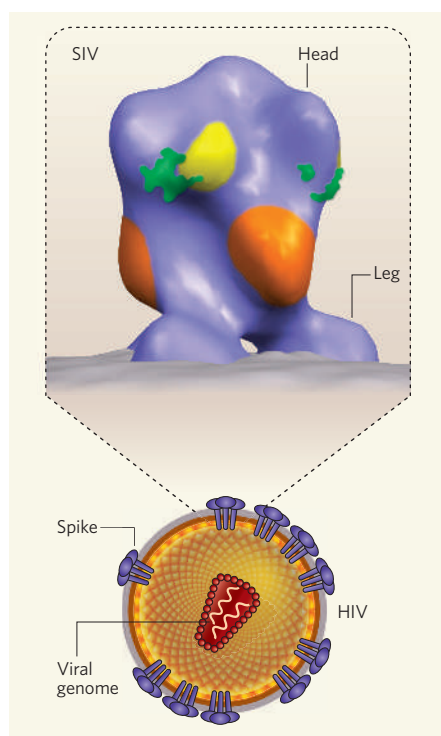


Figure 1 | Envelope spikes on the surface of HIV and SIV. Zhu *et al.*¹ deduce that the spikes on HIV form clusters, and show that a single SIV spike is made of three (gp120–gp41) complexes in a tripod-like assembly, with a head structure made mostly of gp120 sitting on three gp41 legs. The gp41 extends upwards into the head structure. The sugar faces of gp120 are at the top of the spike, and the CD4-binding site (green patches) is close to variable loop structures (red and yellow patches). So the spike displays poorly immunogenic sugars and immunogenic but variable loops, while restricting access to more constant parts of the complex such as the CD4-binding site of gp120 and most of gp41. Targeting the conserved exposed parts of the spike, such as the legs of gp41, using antibodies induced by vaccination, may provide protection against many different HIV strains. (Inset reproduced from ref. 1.)

ture is limited, and it is not clear how many spikes there are on each virus or how they are distributed. The structure of the extracellular part of gp41 has been resolved, but only in a

conformation adopted after the virus connects with the target cell — little is known about what gp41 looks like before this interaction². The structures of isolated truncated gp120 'core' molecules from HIV and SIV have also been solved^{3,4} and have been used to propose models for spikes^{4,5} in which gp120 and gp41 proteins assemble to form a trimer of heterodimers, that is (gp120–gp41)₃. Indeed, electron tomography of fixed and stained preparations of HIV and SIV had revealed trimeric envelope spikes, but many other details were not visible⁶. Now Zhu *et al.*¹ have studied unfixed, unstained, frozen hydrated HIV and SIV using cryoEM tomography, and their low-resolution structure provides valuable information (Fig. 1).

Diagrams tend to present HIV as fairly densely covered in evenly spaced spikes, an image probably based on early electron microscopy studies. Indeed, this is the picture Zhu *et al.* observed for an SIV strain that harbours a mutation enhancing the expression of spikes at the virus surface. These mutant particles displayed 73 ± 25 spikes covering the entire surface. Normal HIV, however, had only 14 ± 7 spikes, and these tended to cluster together.

Such clustering has several implications. First, it challenges a previous proposal that spikes are tethered in an ordered array of pores in the viral matrix underneath the envelope. Second, it suggests that the HIV surface may have yet another feature that evolved to avoid immune detection. Antibody responses to irregular displays of spikes are expected to be less than those elicited by dense, ordered arrays. Third, it is tempting to speculate that clusters of spikes might help the viruses to enter cells. Many researchers have argued that the interaction of multiple spikes with multiple target-cell receptors, namely CD4 and the chemokine receptors CCR5 or CXCR4, is needed to allow formation of a pore in the cell membrane and the transfer of genetic information from virus to target cell. Other researchers, however, have recently argued that just a single HIV spike may be sufficient to allow infection of a cell⁷.

A drawback of the Zhu *et al.* study, and of biophysical studies of HIV generally, should be noted. Under typical experimental conditions, most HIV particles are either not infectious or very slow to infect, and electron microscopy cannot distinguish between infectious and non-infectious particles. It could be that the only virus particles that do effectively infect are those with the highest number of spikes (up to 35 spikes per particle are described by Zhu *et al.*) and/or those particles with particular clustering patterns that are rarely seen.

Nonetheless, Zhu *et al.* provide some stunning images, particularly those of a tripod shape where a head composed mostly of three gp120 molecules balances on three gp41 legs, deduced by averaging data from a large number of SIV spikes. The image is not unlike depictions of menacing aliens that recur in films and books (for a selection try searching Google Image using 'tripods'). The legs are well separated, in contrast to most representations so far. This may help to explain the observation that the external parts of gp41 near the membrane are accessible to neutralizing antibodies, and will encourage vaccine designers to target these regions. The gp120 molecules sit atop the legs, with a sugar-coated face upwards and the variable loops along the

side of the spike, probably restricting antibody access to the crucial CD4-binding site.

What next? Independent corroboration of the tripod structure from different strains and differently treated HIV and SIV preparations is highly desirable. Atomic force microscopy studies have given a different view of the HIV envelope spike⁸, so determining which view is more likely in infectious particles is an important next step. Also, vaccine designers crave a high-resolution structure of an intact native HIV envelope trimer, and this is surely one of the most important unsolved structures in biomedicine. ■

Dennis R. Burton is in the Departments of Immunology and Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
e-mail: burton@scripps.edu

1. Zhu, P. *et al.* *Nature* **441**, 847–852 (2006).
2. Zwick, M. B., Saphire, E. O. & Burton, D. R. *Nature Med.* **10**, 133–134 (2004).
3. Kwong, P. D. *et al.* *Nature* **393**, 648–659 (1998).
4. Chen, B. *et al.* *Nature* **433**, 834–841 (2005).
5. Kwong, P. D. *et al.* *J. Virol.* **74**, 1961–1972 (2000).
6. Zhu, P. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 15812–15817 (2003).
7. Yang, X., Kurteva, S., Ren, X. & Sodroski, J. *J. Virol.* **80**, 4388–4395 (2006).
8. Kuznetsov, Y. G., Victoria, J. G., Robinson, W. E. Jr & McPherson, A. J. *J. Virol.* **77**, 11896–11909 (2003).

MATERIALS SCIENCE

Oxygen breaks into carbon world

Pulickel M. Ajayan and Boris I. Yakobson

When oxygen atoms bind to a graphite surface, they fall into line and make bridges across carbon atoms. This is the spearhead of a chemical attack in which the atomic arrangement of solid carbon is torn apart.

What happens when we burn carbon? Combustion seems such a simple reaction, but at an atomic scale it is the result of several steps. When graphite burns, for example, an intermediate stage is the formation of a graphite oxide¹. This process destabilizes the ordered arrangement of carbon atoms, so that cracks begin to form and the structure breaks up. Controlled oxidation reactions are useful for the preparation of very thin graphite flakes, or for chopping carbon nanotubes into shorter lengths, but the details of how oxygen attacks carbon bonds to break up the atomic structure of graphite have never been understood. Writing in *Physical Review Letters*, Je-Luen Li *et al.*² provide an explanation for this fundamental process.

The carbon atoms in graphite are arranged in a hexagonal lattice like a honeycomb, with each carbon bonded to three others, forming flat sheets held together by weak van der Waals forces. The bonds between the atoms are chemically inert, so unless there are flaws in the lattice, such as missing carbon atoms, the interaction of oxygen gas with a graphite sur-

face is weak. During oxidation, the graphite oxide that forms is a complex structure¹ in which oxygen-containing groups are randomly attached to the honeycomb lattice. These groups, which can take part in varied surface chemistry, mainly attach at defects or edge-atom sites in the lattice, where the carbon atoms are not fully bonded to other atoms. So far so good, but how do the surface-bound oxygen-containing groups trigger the break-up of the carbon lattice, ultimately destroying the entire graphitic structure?

One kind of chemical group that forms on the graphite surface is known as an epoxy bridge, where a single oxygen atom bonds to two adjacent carbon atoms, forming a triangle. In their paper, Li *et al.*² describe how the stress generated by these epoxy bridges leads to unravelling of the graphite lattice. Each epoxy bridge is severely strained, because the geometry of the incorporated carbon atoms has changed. Where the bridged atoms were once only bound to other carbon atoms in a planar, hexagonal honeycomb arrangement, they are

now bound to an oxygen atom sitting above the lattice surface, in place of a lattice carbon atom, and adopt an almost three-dimensional, distorted form. This new geometry doesn't fit well in the remaining lattice — it's like trying to fit a square peg into a round hole.

Mechanistically, the oxygen atom acts as a minuscule wedge, pushing apart the bridge's carbon atoms and stretching the carbon-carbon bond. The epoxy groups do not act individually, but cooperate. The authors' extensive calculations², based on density functional theory, show that side-by-side parallel positioning of the epoxy bridges is energetically favoured, so that they tend to line up on the graphite surface (Fig. 1). As a result, they collectively induce enough tension in the underlying lattice to break the native carbon bonds.

The feedback guiding such organized attachment of oxygen to graphite resembles generic brittle fracture, where the site of the next bond failure is determined by the existing crack as it propagates through the material. Having broken up carbon bonds, the bridging oxygen atoms go on to hold the fractured graphite sheet together, forming a seam between the separated lattice fragments. These fragments are held together at an angle to each other, to comply with the chemically preferred obtuse angle at the oxygen joints. The emerging network of such oxygen-zipped ridges results in the crumpling of a single graphite layer, known as a graphene sheet. This sheet flakes away from the stack of solid graphite, breaking the weak van der Waals forces that once bound it (Fig. 1).

The emerging picture of isolated epoxy groups falling into ranks, to take part in a well-orchestrated serial bond-breaking process, is rather appealing. But the kinetics require further explanation. Randomly bound epoxy groups are unaware of the energetic benefits of forming lines. The epoxy groups can only find such stable alignments as a result of rapid hopping around, but this is not easy to reconcile with the high energy barriers to such hopping. Further study is also required to determine exactly how visible faults or crack discontinuities emerge. Whatever the details may be, the lines of epoxy groups predetermine the tear pattern, like a perforation line directs the tear in a sheet of postage stamps. The resulting fault lines become clearly visible in optical microscope images of oxidized graphite².

A similar sequence of events can also be expected in the oxidation of carbon nanotubes. These share the same bonding as in graphite, except at the end caps where defects are present. Oxidation treatment in strong acids has led to the selective removal of end caps³ and the breaking of long nanotubes into small pieces⁴. One would expect epoxy groups to line up circumferentially in a nanotube, leading to its transverse fracture into shorter segments, in a useful cutting process. On the other hand, uncontrolled oxidation can severely compromise the strength of nanotubes⁵.

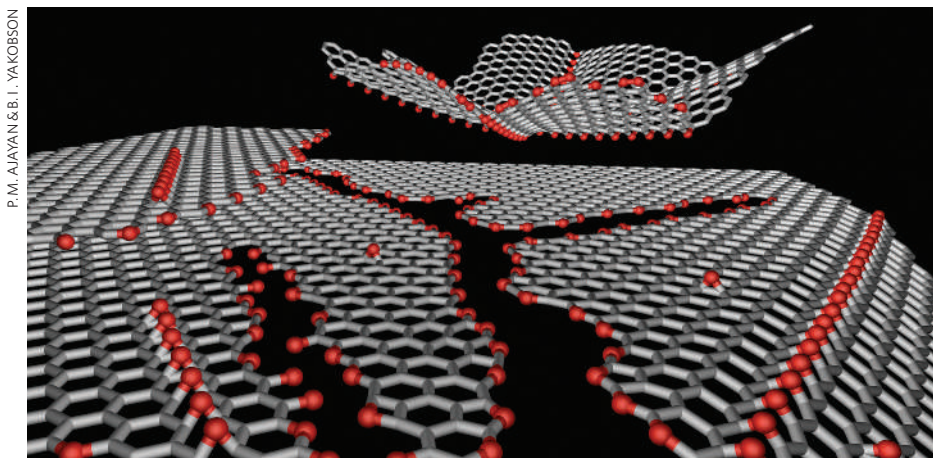


Figure 1 | Oxygen attack on graphite. Li *et al.*² show that when oxygen atoms (in red) bind to a graphite surface, they form two-legged epoxy bridges, which line up to lower their energy. This exerts a collective tension, breaking the underlying carbon-carbon bonds. Structural relaxation around the emerging ridges results in crumpling of the initially flat graphite sheets. This eases their separation, giving rise to distorted flakes of graphite, such as the one at the top of the figure. Faults along the oxygen trails lead to further mechanical fractures.

The paper by Li *et al.*² provides insight into the atomic-level mechanisms of oxidation in carbon. Graphite and its artefacts, such as carbon nanotubes, are materials with a wide range of uses, from lubrication to electronics. Controlled oxidative scission to extract nanoscale graphitic structures (for example, cut-to-size nanotubes⁶ or nanosize graphene sheets⁷) from larger domains of these materials would be an extremely powerful technique for all sorts of applications. Understanding how oxygen breaks up the atomic structure of graphite could lead to a whole new area of nanotechnology based on nanoscale graphite origami⁸.

Pulickel M. Ajayan is in the Department of Materials Science and Engineering, Rensselaer

Polytechnic Institute, Troy, New York 12180, USA.
e-mail: ajayan@rpi.edu

Boris I. Yakobson is in the Department of Mechanical Engineering and Materials Science, and the Department of Chemistry, Rice University, Houston, Texas 77005, USA.
e-mail: biy@rice.edu

1. He, H., Klinowski, J., Forster, M. & Lerf, A. *Chem. Phys. Lett.* **287**, 53–56 (1998).
2. Li, J.-L. *et al. Phys. Rev. Lett.* **96**, 176101 (2006).
3. Ajayan, P. M. & Iijima, S. *Nature* **361**, 333–334 (1993).
4. Liu, J. *et al. Science* **280**, 1253–1257 (1998).
5. Dumitrica, T., Hua, M. & Yakobson, B. I. *Proc. Natl Acad. Sci. USA* **103**, 6105–6109 (2006).
6. Ziegler, K. J. *et al. J. Am. Chem. Soc.* **127**, 1541–1547 (2005).
7. Berger, C. *et al. Science* **312**, 1191–1196 (2006).
8. Ebbesen, T. W. & Hiura, H. *Adv. Mater.* **7**, 582–586 (1995).

NEURODEGENERATION

Good riddance to bad rubbish

Daniel J. Klionsky

Autophagy — cellular ‘self-eating’ — can be induced by stress, but it also acts continuously in a housekeeping role, disposing of unwanted proteins. Can it protect against neurodegenerative diseases?

Alzheimer’s, Parkinson’s and Huntington’s diseases are names we hear with a certain dread. These devastating illnesses, typically associated with ageing, result from the death of neurons. The cause of cell death is not known, but the onset of the disease symptoms is often accompanied by the appearance of large aggregates of particular proteins, such as A β in Alzheimer’s, α -synuclein in Parkinson’s, or huntingtin in Huntington’s disease¹. These are normal proteins — everyone has them — although their function is not always clear. For years, the consensus theory has been that the aggregated

proteins lead directly to cell death. But perhaps the cellular housekeeper that should get rid of the proteins is at fault? There have been hints that autophagy may have a role in protecting against neurodegeneration, based on studies with human cell lines or animals with mutations that predispose them to these diseases¹. In this issue, Komatsu *et al.*² (page 880) and Hara *et al.*³ (page 885) provide the first genetic evidence that the housekeeping role of autophagy is essential for preventing neurodegenerative disease in healthy animals.

Protein aggregates have long been consid-

ered the major culprits in neurodegenerative disease because of the linear correlation between the age of disease onset and the time of aggregate appearance. Furthermore, certain mutated forms of the implicated proteins are more prone to aggregation and/or resistant to degradation, and result in earlier onset of neurological symptoms. Accordingly, many scientists in the field of neurodegeneration have focused on these mutated proteins and the accompanying large aggregates or ‘inclusions’ in cells.

Cells have several mechanisms to dispose of proteins when they misfold, become damaged or are no longer needed. One of the main degradation systems is the proteasome, a multi-subunit enzyme that breaks down proteins that have been tagged with ubiquitin. The proteasome only degrades unfolded, monomeric proteins, however, so it cannot handle protein aggregates. In addition, some of the mutant neuronal proteins are not good substrates for the proteasome.

The other principal degradation system is macroautophagy (which we shall refer to here as autophagy). The hallmark of this process is the formation of double-membrane bubble-like ‘vesicles’ that sequester portions of the cell’s cytoplasm and deliver them to an organelle called the lysosome, where they are broken down (Fig. 1, overleaf). Autophagy can be induced by starvation and various hormonal stimuli⁴. Autophagic vesicles, or autophagosomes, can engulf entire organelles as well as the large aggregates generated by misfolded neuronal proteins. So, there is potential therapeutic value in being able to regulate autophagy to prevent or ameliorate some diseases. Recent data indicate, however, that the large protein aggregates are not the toxic species in these conditions^{5,6}. Rather, the soluble or micro-aggregated forms may be the ones that cause cell death. What, then, is the role of autophagy, and its capacity to sequester large structures, in protecting against neurodegeneration?

Komatsu *et al.*² and Hara *et al.*³ have engineered mice that lack the *Atg7* and *Atg5* autophagy genes, respectively. Both groups have used a genetic trick to delete the gene only from neural cells and only during later stages of embryogenesis, in order to bypass developmental defects that would arise from the elimination of the corresponding gene products constitutively (that is, in all cells throughout development). In both cases, mice lacking the autophagy genes develop symptoms of neurodegeneration, including neuronal cell death.

One reason these studies are of such significance is that they examine mice that are not genetically prone to neurodegenerative disease — the genes encoding the various neuronal proteins in these mice do not have the mutations associated with early onset of disease symptoms. These are healthy animals in which autophagy cannot be acting as an induced cytoprotective response to damaged proteins. This implies that the basal, housekeeping

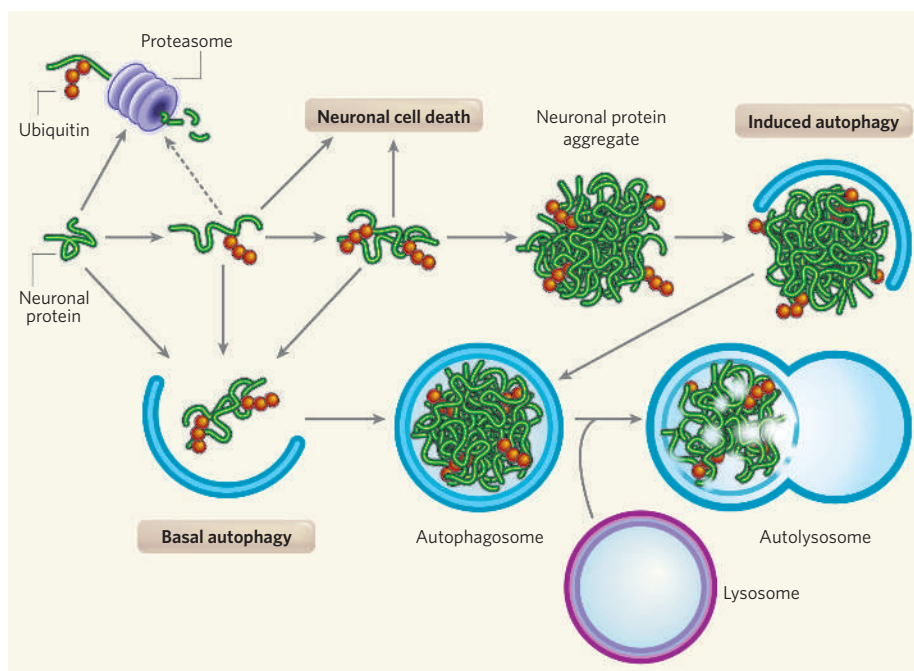


Figure 1 | The role of autophagy in protecting against neuronal cell death. Certain key neuronal proteins may misfold. These misfolded proteins can become ubiquitinated (red circles) and degraded by the proteasome. However, these proteins may be poor substrates for the proteasome, and instead will accumulate in the cytoplasm, making microaggregates and possibly leading to cell death. Basal autophagy can keep the levels of these proteins low enough to prevent toxic effects by sequestering them inside autophagosomes that deliver them to the lysosome for subsequent degradation. The misfolded proteins may also form large aggregates or inclusion bodies that can induce an autophagic response. These large aggregates may be more protective than harmful.

function of autophagy is involved in preventing neurodegeneration.

The previous focus on the role of autophagy in removing large protein aggregates resulted in part from the emphasis on studying patients or animal models with genetic mutations that predisposed the individual to neurodegeneration. In these cases, autophagy might be induced, but its cytoprotective effect could be limited because these large aggregates are not toxic, or are not the initial toxic form. The large aggregates highlighted by studies of mutant proteins may actually be a relatively terminal state of neurodegeneration, rather than the crucial intermediate that leads to disease onset, and where therapeutic intervention may be most effective.

Loss of autophagy in the mice engineered by Komatsu *et al.*² and Hara *et al.*³ also eventually leads to the formation of inclusion bodies analogous to those seen in neurodegeneration. But the initial result of the loss of the *Atg* genes was accumulation of diffuse ubiquitinated proteins in the cytoplasm of the neural cells; the large inclusion bodies were seen later². One implication of this is that everyone, even people without genetic mutations that promote the formation of the large protein aggregates, is potentially susceptible to neurodegenerative disease, and autophagy is a key protective mechanism in otherwise healthy individuals.

As Hara *et al.*³ put it, “a low level of constitutive autophagy must be important for intracellular clearance under normal conditions.”

Even though this seems like an uncontroversial statement, there have previously been few data to support it. Basal autophagy may be particularly vital in neurons, as these cells do not grow and cannot eliminate waste such as protein aggregates by cell division. In addition,

the brain typically does not experience a shortage of nutrients, so autophagy is generally not induced there as part of a starvation response⁷. The current studies therefore suggest that basal autophagy, the routine clearance of the cytosol, is in fact crucial for maintaining healthy cells. These papers will shift the focus of the field of autophagy and neurodegeneration away from large protein aggregates and induced autophagy.

Finally, Komatsu *et al.*² show that the proteasome functions normally in their experimental mice; however, it is obviously not capable of handling all waste disposal for the cell in the absence of autophagy. So why isn't autophagy normally just turned on at higher levels to prevent neurodegeneration prophylactically? Unfortunately, too much autophagy may cause other problems, including autophagic cell death. Routine, but limited, induction of autophagy may have beneficial effects, particularly in individuals who are prone to certain neurodegenerative diseases, but additional studies will be needed to determine whether this is the case.

Daniel J. Klionsky is in the Life Sciences Institute, Department of Molecular, Cellular and Developmental Biology, and the Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA.
e-mail: klionsky@umich.edu

1. Rubinsztein, D. C. *et al.* *Autophagy* **1**, 11–22 (2005).
2. Komatsu, M. *et al.* *Nature* **441**, 880–884 (2006).
3. Hara, T. *et al.* *Nature* **441**, 885–889 (2006).
4. Levine, B. & Klionsky, D. J. *Dev. Cell* **6**, 463–477 (2004).
5. Arrasate, M., Mitra, S., Schweitzer, E. S., Segal, M. R. & Finkbeiner, S. *Nature* **431**, 805–810 (2004).
6. Tanaka, M. *et al.* *J. Biol. Chem.* **279**, 4625–4631 (2004).
7. Mizushima, N., Yamamoto, A., Matsui, M., Yoshimori, T. & Ohsumi, Y. *Mol. Biol. Cell* **15**, 1101–1111 (2004).

QUANTUM PHYSICS

United through repulsion

Leonardo Fallani and Massimo Inguscio

Mutually repulsive atoms placed at periodic intervals in a ‘crystal of light’ can, counterintuitively, be forced into stable couplings. That theoretical prediction has just seen experimental confirmation.

Nature likes to save energy. Whenever independent particles can arrange themselves freely, they choose the configuration that minimizes their total energy. That's why a molecule can be formed from two or more atoms: the bound configuration is stable because the total energy of the atoms is smaller when they are close together than when they are far apart. On page 853 of this issue, Winkler *et al.*¹ report a fundamental extension to this concept of a molecule — atom pairs bound together through repulsive, rather than attractive, forces.

It seems intuitive that, to obtain a bound system of particles, an attractive force between

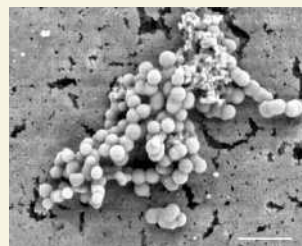
them is required. But it is also easy to see that attraction by itself is not sufficient: to reach an equilibrium configuration, a repulsive force must also be present, or the system would simply collapse. This phenomenon is quite general. In astrophysics, for example, certain types of supernovae are the results of implosions that occur when a star's self-gravity is no longer balanced by the pressure generated by thermonuclear reactions in its core. Stable stars, on the other hand, exist when some repulsive force counterbalances the attraction of gravity. In a neutron star, this repulsion is provided by the rules of quantum mechanics, which forbid

BIOMATERIALS

Silk spin-off

Evolution has produced some remarkable biological materials, and materials scientists try to cherry-pick the best properties to produce 'biomimetic' composites that transcend the originals. To this end, David Kaplan and colleagues have created a fusion protein that combines spider silk protein and a peptide (R5) that helps to deposit the silicate skeletons of microorganisms called diatoms (*Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0601096103). In diatom biosilicates, the organic

protein component improves the toughness of the inorganic silicon, and allows fine control of the microstructure — giving complex skeletons that are produced under mild aqueous conditions. Spider silk, on the other hand, is renowned for its tensile strength and, of particular importance for this study, can assemble itself from its constituent units. Silk assembly can be manipulated experimentally to produce many structures, from the threads seen in nature, to films, gels



and porous matrices. Silk studded with cell-binding domains or other biologically active molecules can even be made, to encourage the formation of tissues such as bone or cartilage.

Kaplan and colleagues cast their silk-R5 fusion protein into films or spun it into fibres, depositing silica

by adding a silicon solution during or after the process. They obtained spherical silica structures 0.5–2 μm in diameter (the picture shows silica deposited on a silk-R5 film; scale bar 2 μm). The strength and resilience of the silk-silica structures have not been tested, but the work shows in principle that they can be produced in different forms, for instance as silk fibres coated entirely in silica. The authors hope that their fusion strategy could produce a variety of composite materials. By adjusting the proteins involved, it might be possible to direct the deposition of titanium dioxide rather than silica, say.

Helen Dell

particles of half-integer spin, such as neutrons, getting too close to one another. Similar quantum-mechanical effects can be reproduced, at very low temperatures only billionths of a degree from absolute zero, on the entirely different scale of atomic physics in the matter phase known as a Bose–Einstein condensate^{2,3}.

What is surprising is that, in the quantum world, attraction isn't even necessary to form a stable bound system: repulsion can be enough. This prediction, like so much of quantum mechanics, is counterintuitive, as one would expect two repelling particles simply to fall apart to minimize their interaction energy. But in the presence of a periodic spatial perturbation this should no longer be true. Here, the energy of a particle cannot vary continuously, but is restricted to particular ranges of values. A pair of two repelling particles can therefore be stable simply because, if it fell apart, conservation of energy would require the two isolated atoms to have energies that would fall in a forbidden range.

Although periodic structures are the natural forms in which the atoms arrange themselves into crystals — and as such are very common in physics — interactions with the environment in the solid state dissipate energy, preventing the observation of the bound, repulsive pairs predicted by the theoretical model. Winkler and colleagues¹ circumvent this obstacle by using a periodic structure made of light. In such an 'optical lattice', ultracold atoms are ordered in a regular array of microscopic traps caused by the interference of two or more laser beams. The resulting ordered system of atoms resembles a solid-state configuration, and neutral atoms in an optical lattice do indeed share many properties with electrons in a metal. In contrast to a solid-state crystal, however, light crystals are free from defects and dissipative vibrations, and have thus proved ideal systems for the investigation of otherwise elusive quantum-physical phenomena⁴.

In Winkler and colleagues' experiments, the capabilities of the optical lattice are combined with a powerful tool that atomic physicists can

now use to change the nature of the interactions between two atoms, transforming attractions into repulsions, and vice versa. This trick is achieved by placing cold atoms in a uniform magnetic field, and tuning the intensity of the field across a resonance — known as a Feshbach resonance⁵ — that occurs when the energy of a bound molecular state is exactly equal to the energy of two colliding atoms. Following on from the observation of Bose–Einstein condensation^{6,7} (which involves 'bosonic' atoms of integer spin; atoms of half-integer spin are termed 'fermionic'), Feshbach resonances have made possible the observation of striking phenomena — including the formation of cold molecules⁸ and culminating in the observation of superfluidity among ultracold fermionic atoms⁹.

Winkler *et al.* take a cold gas of bosonic rubidium (⁸⁷Rb) atoms, which naturally have repulsive interactions, and first fill a three-dimensional, cubic optical lattice such that each lattice site confines either no atoms, or two identical atoms (Fig. 1). The authors then use a Feshbach resonance to control the interaction between the atoms forming the pairs. When interactions are cancelled, the pairs are not stable and the atoms quickly fall apart, diffusing within a few milliseconds across the lattice and hopping from one site to the next.

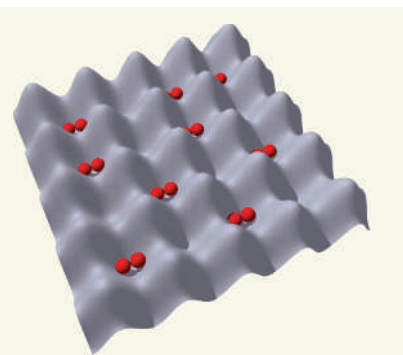


Figure 1 | Repulsive pairs. Repulsion between pairs of identical rubidium atoms in an artificial optical lattice causes them to be stable against separation, as Winkler and colleagues¹ demonstrate.

When repulsive interactions are restored, however, the separation process is halted. The pairs of atoms then live much longer together — several hundred milliseconds — clearly demonstrating that their stability is induced by the mutual repulsion of the constituents. The authors obtain additional information about the nature of the pairs by measuring the velocity distribution of the atoms and the energy of the bond.

This new type of bound object remains stable because, in the structured environment of the lattice, the large repulsive interaction between the atoms cannot be converted into kinetic energy. This result is general, as magnetic fields could be used to tune the strength of the interaction between other atomic species. An even broader impact could be achieved by extending the concept to mixtures of different atoms, and also by playing with quantum gases of different nature, both bosonic and fermionic⁴. Once more, it seems, optical lattices are demonstrating how useful they are for investigating many-body phenomena unobservable in other systems in which the interaction with the environment is too strong. As cold atoms are themselves a type of experimental simulator of quantum-physical effects, the result is a toolbox¹⁰ that kits us out to explore fascinating new avenues in quantum information and quantum states of matter.

Leonardo Fallani and Massimo Inguscio are at the European Laboratory for Non-Linear Spectroscopy (LENS), Università degli Studi di Firenze, Polo Scientifico, Via Nello Carrara 1, I-50019 Sesto Fiorentino, Italy.
e-mail: inguscio@lens.unifi.it

1. Winkler, K. *et al.* *Nature* **441**, 853–856 (2006).
2. Donley, E. A. *et al.* *Nature* **412**, 295–299 (2001).
3. Anglin, J. R. & Ketterle, W. *Nature* **416**, 211–218 (2002).
4. Bloch, I. *Nature Phys.* **1**, 23–30 (2005).
5. Feshbach, H. *Ann. Phys.* **5**, 357 (1958).
6. Cornell, E. A. & Wieman, C. E. *Rev. Mod. Phys.* **74**, 875–893 (2002).
7. Ketterle, W. *Rev. Mod. Phys.* **74**, 1131–1151 (2002).
8. Donley, E. A. *et al.* *Nature* **417**, 529–533 (2002).
9. Grimm, G. *Nature* **435**, 1035–1036 (2005).
10. Jaksch, D. & Zoller, P. *Ann. Phys.* **315**, 52–79 (2005).

NEUROECONOMICS

Best to go with what you know?

Daeyeol Lee

In a changing world, how do we decide our best option? How do we settle between picking something familiar or trying out a new, possibly more rewarding, choice?

You step into a music store: how do you choose which CD to buy? If you pick something by your favourite composer, Schubert say, you know you will enjoy it. But you might want to expand your repertoire by trying a less familiar piece, for example Mahler's symphony No. 10. You might not like it; on the other hand, you might discover a new favourite. So how do you decide between sticking with what you know and like, and trying something more adventurous that might prove even more rewarding? On page 876 of this issue, Daw *et al.*¹ provide valuable insight into how people balance the two basic impulses of exploiting what they know and exploring new options.

Standard economic theories tend to assume that people always make the decision that gives them the greatest reward — in terms of its subjective value or utility to the person. In reality, however, people often do not know which option will produce the best outcome, as even familiar environments change continually, so decision-making strategies must be adjusted through experience². In reinforcement-learning theories³, estimates of future rewards expected from alternative actions are referred to as their value functions, and these must be updated appropriately to reflect the actual rewards. Choosing the action with the maximum value function is referred to as exploitation. Because value functions are only estimates of future rewards, however, it is often desirable to try out actions that might seem to have suboptimal value functions, and this is known as exploration.

To see how people deal with making decisions in changing circumstances, Daw *et al.*¹ asked their subjects to play a computer game where they could choose among four slot machines. Each machine was assigned a different payout level, and each time the subject played, the payout they received from their chosen machine was determined randomly on the basis of the mean and the standard deviation of the assigned payout. In addition, the mean payout of each slot machine drifted randomly over time, preventing the subjects from ever knowing accurately which slot machine would reward them the most.

Daw *et al.* considered the ways in which people might resolve the exploration–exploitation dilemma. One possibility, known as the ϵ -greedy rule, is to select the action with the maximum value function most of the time, but to choose randomly among the remaining

options with a small probability (ϵ). Alternatively, people might explore more frequently if the differences among value functions of the various options are relatively small. This behaviour is described by the 'softmax' rule, which states that the probability of choosing a particular action is given by the Boltzmann distribution of the value functions (analogous to the Boltzmann distribution in physics that describes the behaviour of particles according to their energies).

To illustrate the difference between these two rules, let us consider the choice of Schubert versus Mahler, and imagine that the

value function is higher for Schubert than for Mahler (shown by the light-blue area in Fig. 1). If there is no exploration, one would always choose Schubert, consistent with the utility maximization posited by standard economic theories. According to the ϵ -greedy rule, Mahler will be chosen occasionally (with probability ϵ). According to the softmax rule, the probability that less-preferred Mahler will be chosen depends on the difference in the value functions of these two composers, and increases as Mahler's value function becomes more similar to Schubert's.

The results of Daw and colleagues showed that in their slot-machine game the softmax rule accounted for the pattern of exploration better than the ϵ -greedy rule. Yet another strategy of exploration in decision-making is to increase the probability of choosing a particular action according to the uncertainty of its outcome. Daw *et al.* found that applying this so-called 'exploration bonus' did not account for the data any better than the simple softmax rule. So it seems that the subjects did indeed explore according to the softmax rule.

To understand how the brain decides to explore, Daw *et al.* examined the activity in the subjects' brains during the slot-machine game using functional magnetic resonance imaging. Consistent with findings from previous studies, they found that several brain areas, such as the orbitofrontal cortex, displayed changes in activity that were related to the value functions of the chosen action². More notably, having classified a subject's choice in each trial as exploitation or exploration, the authors looked for brain areas that showed increased activity associated with each choice. They did not find any areas that became more active during exploitation, but they did find that exploration was accompanied by stronger activation in the frontopolar cortex, a region considered important for the control of cognitive functions⁴.

The findings of Daw *et al.* not only provide a unique insight into how we make decisions, but also raise many exciting questions for future research. For example, the randomness of the choice generated by the softmax rule can be specified by a parameter analogous to the temperature in the Boltzmann distribution. At a high 'temperature', all actions are chosen more or less randomly, whereas at a low 'temperature' preferred actions are chosen deterministically. Is it possible that the temperature parameter in the softmax rule decreases, and decision-makers explore less, as they get more familiar with a particular task? If so, what are the underlying neural mechanisms? Changes in the parameters of learning models, such as temperature, are known as meta-learning^{5,6}, and this process may be controlled by neuromodulators — chemicals such as noradrenaline and acetylcholine that affect the efficacy of other neurotransmitters^{7,8}. But the precise roles of these neuromodulators in reinforcement learning and decision-making have still to be defined. As beautifully exemplified

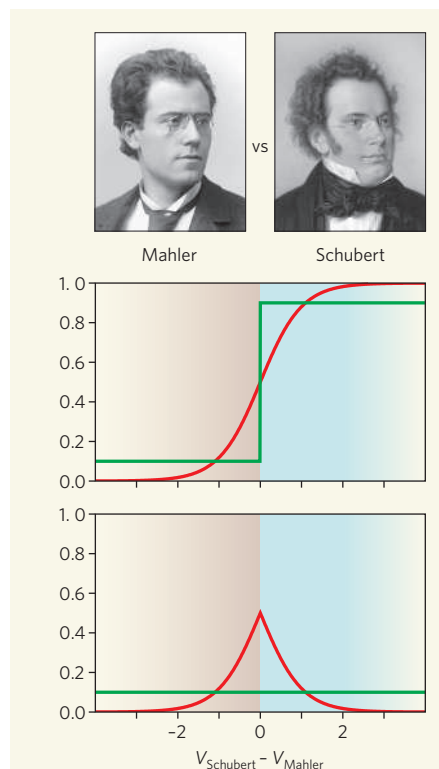


Figure 1 | The contribution of exploration to making a decision. The probability (P) of choosing a particular composer (for example, Schubert rather than Mahler) can be determined by the difference in the value functions between the two composers ($V_{\text{Schubert}} - V_{\text{Mahler}}$). According to the ϵ -greedy rule (green line), the choice to explore a less-favoured option is made randomly with some probability (ϵ ; here $\epsilon = 0.1$). By contrast, according to the softmax rule (red line), the probability of exploration increases gradually as the difference in the value functions approaches zero. Daw *et al.*¹ show that in their slot-machine game, subjects seemed to explore options according to the softmax rule.

LEBRECHT MUSIC & ARTS PHOTO LIBRARY/ALAMY; AKG-IMAGES/E. LESSING

by Daw *et al.*¹, exploring these questions will require interdisciplinary approaches, in which the computational theories inspire and guide behavioural and neurobiological experiments. ■

Daeyeol Lee is in the Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, USA.
e-mail: dlee@cvs.rochester.edu

1. Daw, N. D. *et al.* *Nature* **441**, 876–879 (2006).
2. Lee, D. *Curr. Opin. Neurobiol.* **16**, 191–198 (2006).
3. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
4. Ramnani, N. & Owen, A. M. *Nature Rev. Neurosci.* **5**, 184–194 (2004).
5. Schweighofer, N. & Doya, K. *Neural Netw.* **16**, 5–9 (2003).
6. Soltani, A., Lee, D. & Wang, X.-J. *Neural Netw.* (in the press).
7. Doya, K. *Neural Netw.* **15**, 495–506 (2002).
8. Aston-Jones, G. & Cohen, J. D. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).

SOLID-STATE CHEMISTRY

A glass of carbon dioxide

Paul F. McMillan

Carbon is unusual in its family of elements because it has gaseous oxides. But under high pressure, carbon dioxide forms crystalline solids and can become a glass — so revealing the chemical family resemblance.

Everyone is familiar with the common forms of silicon dioxide (SiO₂, silica), such as the crystalline version known as quartz, the major component of sand. When melted and cooled rapidly, and usually mixed with metal oxides, sand forms silica-based glasses that we use to make many useful things, ranging from windows to champagne bottles. A glassy dioxide of germanium (GeO₂, germania) is also known, and is added to the kilometre-long silica glassy fibres used in optical telecommunications, to control their refractive index. Carbon belongs to the same family of elements as silicon and germanium, but carbon dioxide (CO₂, carbonia) glass has remained a theoretical possibility. On page 857 of this issue, however, Santoro and colleagues¹ show that carbonia glass can be made when a high-density form of solid CO₂ is melted and cooled under high-pressure conditions.

Carbon, silicon and germanium are the first three members of group IV of the periodic table, which also includes the heavier elements tin and lead. These last two elements form both monoxides and dioxides, which have been used as pigments since ancient times². In contrast, only the dioxides of silicon and germanium occur as stable solids. Once we reach carbon, the lightest member of the series, we find that both carbon monoxide and carbon dioxide are stable molecules, but these exist as gases at ambient temperature and pressure. As we rise up a group of the periodic table, such anomalous jumps in the properties of compounds formed from the elements of that group often occur. Smooth changes in chemical and physical properties are generally observed among the heavier members of the group, but the lightest element behaves quite differently.

Carbon dioxide is a linear molecule with strong carbon–oxygen double bonds. Freezing or high-pressure treatment condenses this gas into solid forms, in which the molecules remain intact, even under extreme pressuriza-

tion to 80 GPa at ambient temperature³. Recently, however, experiments showed that under simultaneous high-pressure and high-temperature conditions, CO₂ molecules undergo bond-breaking and re-formation reactions, producing a three-dimensional network of polymerized tetrahedral CO₄ units. This network is analogous to the crystalline silica structures found in minerals such as cristobalite, tridymite or quartz^{4,5}, although it reverts to solids containing CO₂ molecules at pressures below 1 GPa. Other nitrogen- and oxygen-containing molecules also undergo solid-state chemistry at high pressure⁶. These results have generated great excitement among chemists.

The newly prepared materials might have useful properties for technological applications. Studies on the polymeric, silica-like form of CO₂ suggest that it is 'super-hard' and that it is an optically nonlinear material — for example, it causes doubling of the frequency of light from a laser^{4,5}. The discovery also has implications for geochemistry, because the conditions found in Earth's mantle could induce the formation of these newly discovered forms of carbon oxides. Incorporation of CO₂ into solid-state compounds formed under high temperature and pressure may even lead to methods for disposal of this environmentally problematic gas.

So, a crystalline silica-like form of CO₂ has been made, but how about a glassy version? Enter Santoro and colleagues¹, who have prepared a dense, polymeric amorphous form of this compound that is analogous to silica- and germania-based glasses. They did this by compressing molecular solid CO₂ to pressures of 40–65 GPa with heating, then cooling it to ambient temperature. Their result is at least as exciting as the discovery of the polymeric crystalline carbon dioxide solid, and the chemistry of carbonia-based glasses is now open for exploration.

So far, glassy carbonia is like its crystalline equivalent in that it is not yet recoverable to

ambient conditions, preventing further study of its physical properties. During decompression from the conditions under which it is formed, the glass transforms back to an amorphous version of molecular solid CO₂. Developing carbonia glasses that can be recovered to ambient conditions will be one of the first challenges for future research.

These results¹ have implications for liquid-state physics, and inform us about the conditions under which CO₂ undergoes various phase transitions, including melting⁷. There has been much discussion of the newly recognized phenomenon of polyamorphism — the ability of a material to exist in several different amorphous forms — and of phase transitions between distinct liquid states of a single substance. These are driven by density or entropy changes between glassy or liquid forms at constant chemical composition⁸. It has been suggested that such transitions occur for many liquids and glasses, including water, silica and germania. The observation by Santoro *et al.*¹ that, upon decompression, glassy carbonia transforms from a dense amorphous solid containing networks of single bonds, to an amorphous solid containing molecular CO₂, represents a dramatic example of polyamorphic behaviour, analogous to that observed for amorphous silicon or liquid phosphorus^{9,10}. It follows that anomalies are to be expected in the melting behaviour of crystalline CO₂ solids, or that a phase transition may occur in the liquid state at high pressure, between the molecular and the polymeric forms.

The discovery of mineral-like polymeric and ionic solids that form at high pressure, based on the light elements carbon, oxygen and nitrogen, opens up a new area in solid-state chemistry. Carbonia-based minerals and glasses could give rise to useful technological materials, if we can recover them to ambient conditions. These findings will also help set the rules for understanding structure, bonding and thermodynamic properties as we move our experiments into the high-pressure, high-temperature conditions mimicking those deep inside planetary interiors. ■

Paul F. McMillan is in the Department of Chemistry, Christopher Ingold Laboratory and Materials Chemistry Centre, University College London, 20 Gordon Street, London WC1H 0AJ, UK, and in the Davy–Faraday Laboratory, Royal Institution of Great Britain, London.
e-mail: p.f.mcmillan@ucl.ac.uk

1. Santoro, M. *et al.* *Nature* **441**, 857–860 (2006).
2. Ball, P. *Bright Earth: The Invention of Colour* (Viking, London, 2001).
3. Liu, L.-G. & Bassett, W. A. *Elements, Oxides, Silicates: High-Pressure Phases with Implications for the Earth's Interior* (Oxford Univ. Press, 1986).
4. Iota, V. *et al.* *Science* **283**, 1510–1513 (1999).
5. Yoo, C.-S. *et al.* *Phys. Rev. Lett.* **83**, 5527–5530 (1999).
6. Ermetts, M. I. *et al.* *J. Chem. Phys.* **121**, 11296–11300 (2004).
7. Tschauer, O. *et al.* *J. Phys. Rev. Lett.* **87**, 075701 (2001).
8. McMillan, P. F. *J. Mat. Chem.* **14**, 1506–1512 (2004).
9. Katayama, Y. *et al.* *Nature* **403**, 170–173 (2000).
10. McMillan, P. F., Wilson, M., Daisenberger, D. & Machon, D. *Nature Mater.* **4**, 680–684 (2005).

OBITUARY

Bruce Merrifield (1921–2006)

Inventor of solid-phase peptide synthesis.

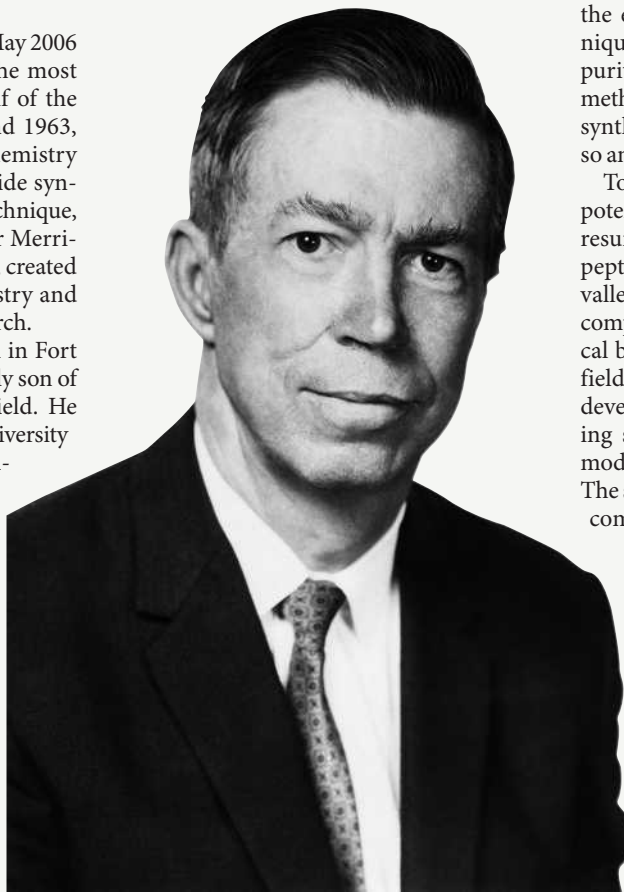
BETTMANN/CORBIS

The death of Bruce Merrifield on 14 May 2006 brings to a close the life of one of the most original scientists of the second half of the twentieth century. Between 1959 and 1963, Merrifield revolutionized organic chemistry by his invention of solid-phase peptide synthesis. This “simple and ingenious” technique, as it was described in the citation for Merrifield’s 1984 Nobel Prize in Chemistry, created a paradigm shift in synthetic chemistry and profoundly affected biomedical research.

Robert Bruce Merrifield was born in Fort Worth, Texas, on 15 July 1921, the only son of George and Lorene (Lucas) Merrifield. He obtained his PhD in 1949 from the University of California, Los Angeles. Immediately after graduating, Merrifield joined the Rockefeller Institute for Medical Research in New York, where he spent his entire career. He became professor in 1966, and was elected to the US National Academy of Sciences in 1972.

At the Rockefeller Institute, Merrifield worked with D. Wayne Woolley on nucleotide biochemistry, and then on the identification and chemical synthesis of putative peptide growth factors. This was the period shortly after the Second World War, when chemists worldwide were devising methods for the preparation of biologically active peptides. These molecules were important targets for organic synthesis, because they constituted a new class of natural product. Synthetic techniques had improved rapidly, but were still rather cumbersome. Using the available chemistries, Merrifield prepared a series of synthetic peptides containing up to seven amino acids, a process that took five years of laborious research. On 26 May 1959, Merrifield wrote in his laboratory notebook: “There is a need for a rapid, quantitative, automatic method for the synthesis of long chain peptides.” He then outlined the principles of a radically new approach to chemical peptide synthesis. Four years later, Merrifield published a seminal paper, entitled “Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide” in the *Journal of the American Chemical Society*.

The solid-phase method consisted of chemically attaching the last amino acid of the target peptide to a solid support, then carrying out all the chemical steps necessary to build the peptide’s amino-acid sequence, and finally releasing the product from the support. As Merrifield envisaged it, solid-phase synthesis



would greatly facilitate the preparation of peptides by enabling the recovery of support-bound intermediate products by simple filtration and washing; in turn, this would enable the use of excess reactants at each step, giving rapid reactions and high yields of the final peptide product.

From the start, Merrifield had also envisaged the automation of chemical synthesis. By 1965 he had implemented this idea in collaboration with John Stewart, a faculty colleague. With Nils Jernberg of the machine shop at the Rockefeller University (as it was known by then), they built an automated synthesizer that could prepare peptides 20 times faster than solution methods, and that allowed the preparation of longer peptide chains.

Solid-phase peptide synthesis was quickly adopted throughout the world, and was used to make numerous analogues of biologically active peptides. Merrifield and his young colleagues concentrated on developing the technique and extending it to new targets. But as solid-phase synthesis was applied to ever-larger molecules, established peptide researchers harshly criticized both Merrifield and his method. It was felt, as a matter of prin-

ciple, that a technique that prevented the full characterization of synthetic intermediates could not give rise to authentic peptide products. The objections of traditional chemists were ultimately refuted by the systematic development of improved solid-phase synthetic chemistry in Merrifield’s and other laboratories during the 1970s and 1980s, and by the emergence of powerful analytical techniques for determining the structure and purity of synthetic peptides. The solid-phase method is now used routinely for the chemical synthesis of peptides and proteins up to 50 or so amino acids in length.

Today, the discovery of vast numbers of potent peptide natural products is leading to a resurgence in pharmaceutical research on peptides. Solid-phase synthesis is the unrivalled technique for the preparation of such compounds in order to determine the chemical basis of their biological function. Merrifield’s ingenious concept led directly to the development of efficient methods for preparing synthetic nucleic acids — essential for modern molecular biology and biotechnology. The solid-phase principle is also at the heart of combinatorial chemistry, which enables the simultaneous preparation of thousands of compounds, and which has changed the fabric of medicinal chemistry and drug discovery since the early 1990s.

Bruce Merrifield is survived by Libby, his wife of 56 years, their six children and sixteen grandchildren. He was a devoted family man who loved hiking and camping with his family. In the last 45 years of his life, Merrifield was afflicted with a progressive skin cancer — apparently the result of radiation treatment for acne in his youth. A man of extraordinary resilience, Merrifield bore this devastating illness with grace and fortitude. In his later years, he nursed his wife back to health after she suffered a serious stroke.

In the 1960s and 1970s, the Merrifield laboratory was a hotbed of new peptide science. Those who were fortunate enough to work with Bruce came to know a man who was devoted to the Rockefeller University, to his research group, and to intellectually rigorous science. His unique approach to his discipline was recounted in an excellent autobiographical memoir, *Life During a Golden Age of Peptide Chemistry* (American Chemical Society, 1993). By creating a completely new approach to organic synthesis, Merrifield pioneered the effective application of synthetic chemistry to the study of biological molecules. His scientific creativity has been equalled by few and rarely surpassed. ■

Stephen Kent

Stephen Kent is in the Department of Biochemistry & Molecular Biology, and the Department of Chemistry, University of Chicago, 929 East 57 Street, Chicago, Illinois 60637, USA.
e-mail: skent@uchicago.edu

Accretion of the Earth and segregation of its core

Bernard J. Wood¹, Michael J. Walter² & Jonathan Wade²

The Earth took 30–40 million years to accrete from smaller ‘planetesimals’. Many of these planetesimals had metallic iron cores and during growth of the Earth this metal re-equilibrated with the Earth’s silicate mantle, extracting siderophile (‘iron-loving’) elements into the Earth’s iron-rich core. The current composition of the mantle indicates that much of the re-equilibration took place in a deep (>400 km) molten silicate layer, or ‘magma ocean’, and that conditions became more oxidizing with time as the Earth grew. The high-pressure nature of the core-forming process led to the Earth’s core being richer in low-atomic-number elements, notably silicon and possibly oxygen, than the cores of the smaller planetesimal building blocks.

From observations of newly formed stars it appears that the Sun’s planetary system formed from a flattened disk of dust and gas which accreted rapidly ($\sim 10^4$ yr) into ‘planetesimals’ ~ 10 km in diameter¹. Gravitational interactions and collisions between these bodies generated Moon-to-Mars-sized planetary embryos in 10^5 – 10^6 yr and planetary bodies on a timescale of 10–100 million years (Myr)¹. Tungsten isotope anomalies in meteorites² demonstrate that their asteroidal parents segregated cores within a few million years of the origin of the Solar System. Applied to the Earth, the same techniques point to protracted (30–40 Myr) metal re-equilibration during planetary growth. Comparison of the composition of the Earth’s mantle with that of presumed protoplanetary material³ indicates high-pressure, high-temperature conditions of core segregation^{4–6}, consistent with the existence of a deep silicate ‘magma ocean’ during much of the Earth’s growth. Once the Earth had attained the size of Mars, however, crystallization of silicate perovskite in the lower mantle led to increasingly oxidized conditions^{6,7}, which ultimately halted the segregation of metal and may have led to a latest phase of sulphide addition to the core^{8,9}. This drawn-out process generated a core containing, in addition to the major components Fe ($\sim 85\%$) and Ni (5%), 1.9% S, 4–5% Si and possibly $>1\%$ O.

Accretion and core formation in planetary objects

The planets of the Solar System originated as dust and gas in the young Sun’s protoplanetary disk. The mechanisms of initial growth towards large bodies are poorly understood, but whether by gravitational instability or simple ‘sticking together’ of aggregates, the process must have formed a large number of 10-km-sized objects rapidly if the condensed materials were to avoid being dragged into the Sun^{10–12}. Once bodies reached this critical size, gravitational perturbation became the dominant mechanism for further accretion through collision. Although some planetesimals were destroyed in collisions, others would have continued to accrete, provided impact velocities were low enough, or impact angles shallow enough, for gravity to recombine impact fragments and ejecta^{1,13–15}. Models show that once a population of larger planetary embryos emerged, they accreted rapidly at the expense of smaller objects in a period of ‘runaway growth’ driven by dynamical friction and gravitational

focusing¹. Accretion eventually became self-regulated by gravitational interactions among planetary embryos and surviving planetesimals in a period of oligarchic growth¹⁶. The few tens of planetary embryos that survived this accretionary phase of about one million years were Moon-to-Mars-sized objects, with masses of the order of 0.01–0.1 times the mass of the Earth¹⁷.

The evidence for early core formation in planetary embryos and planetesimals (radius > 30 km) is substantial. Many meteorites are samples from the asteroid belt between Mars and Jupiter, and are probably a remnant of the early days of accretion. The occurrence of distinct populations of iron and achondritic silicate meteorites attests to early differentiation in planetesimals that were subsequently destroyed during collisions¹⁸. Tungsten (W) isotopes confirm early core formation and place strict time limits on the timing of segregation of metal from silicate¹⁹. ^{182}Hf decays to ^{182}W with a half-life of about nine million years. Hafnium (Hf) is lithophile, having a pronounced preference for silicate phases over metal, whereas W is siderophile, preferring the metal phase. So when a metal core segregates from silicate, the Hf/W ratio becomes elevated in the silicate but is near zero in the metal. If metal segregation occurred while ^{182}Hf was ‘alive’ (<45 Myr after the origin of the Solar System), then silicate reservoirs exhibit positive ^{182}W anomalies and metal reservoirs exhibit negative anomalies relative to undifferentiated proto-nebular material such as the chondritic meteorites. The ‘depleted’ ^{182}W isotopic composition of iron meteorites relative to the chondrite standard constrains metal segregation in precursor objects to within a few million years (<5 Myr) of the origin of the Solar System^{2,19}. This time frame is consistent with statistical models in terms of how long it takes to grow objects large enough (>30 km) to retain their heat¹⁵. Heat retention is important because metal does not segregate from silicate unless it is molten and efficient segregation also requires melting of the silicate^{20,21}.

Several heat sources contributed to melting of protoplanets in the young Solar System. Decay of the short-lived radionuclides ^{26}Al (half-life $t_{1/2} \approx 7 \times 10^5$ yr) and ^{60}Fe ($t_{1/2} \approx 3 \times 10^5$ yr) could have supplied sufficient heat to completely melt an early formed planetesimal^{21–24}. Even in the absence of radioactive decay, the kinetic energy supplied by impact among small, porous planetesimals, and especially from collisions involving larger objects, would have

¹Macquarie University, Department of Earth and Planetary Sciences, North Ryde, New South Wales 2109, Australia. ²University of Bristol, Department of Earth Sciences, Queen’s Road, Bristol BS8 1RJ, UK.

caused substantial melting. Surviving undifferentiated bodies must, therefore, have grown relatively slowly, enabling heat to be radiated away and after substantial decay of the most important radio-nuclides. In small, differentiated bodies, the pressure at which metal and silicate material equilibrated during core segregation would generally have been less than 2 GPa. Equilibration temperature is more difficult to define because young objects can reach very high temperatures from the heating processes discussed above and from potential energy released during settling of metal²¹. Metal–silicate equilibration in early-formed objects may therefore have occurred at a wide range of high temperatures (for example, 1,500 to >3,000 K).

By the end of the oligarchic growth stage the tens of surviving planetary embryos were no longer constrained in well-regulated orbits and began to interact gravitationally, setting up a final, cataclysmic stage of accretion by collision which lasted $\sim 10^7$ years^{1,25}. Many of these objects had proto-cores of iron alloy. A large terrestrial planet like the Earth probably sustained a number of big collisions during accretion, and a late-stage giant impact between a Mars-sized object ($\sim 10^{26}$ g), sometimes referred to as 'Theia', and the proto-Earth ($\sim 10^{27}$ g) is the prevailing theory for the formation of the Moon. This would have provided sufficient energy to melt the proto-Earth completely.

In summary, the Earth accreted over a period of at least 10^7 years from smaller bodies, most of which had already-segregated metallic cores. The energies of impact and metal separation provided sufficient heat to induce substantial melting episodically throughout the accretion and core-formation process. It is in the context of accretion, melting and metal segregation that the process of core formation needs to be viewed.

Chemical signature of core formation

Growth of the Earth from planetary embryos and planetesimals resulted in the substantial partitioning of siderophile elements into the metallic core, leaving lithophile elements behind in the silicate mantle. Observed metal–silicate partitioning behaviour can, in principle, be used to understand the core separation process provided chemical compositions of the bulk silicate Earth (otherwise known as the primitive mantle), the core and the bulk Earth are

known. Primitive mantle is estimated from analyses of mantle peridotites^{3,26}, bulk Earth from the compositions of undifferentiated protoplanetary material, represented by the CI carbonaceous chondrite meteorites and the core is obtained by calculating the difference.

Figure 1 shows a plot of the ratios of elemental concentrations in the bulk silicate Earth divided by those in CI chondrites, normalized to $\frac{[\text{Mg}]_{\text{Earth}}}{[\text{Mg}]_{\text{CI}}} = 1.0$ to correct for the high volatile contents of the meteorites. The elemental ratios are plotted as a function of the temperature by which 50% of the element of interest would have condensed during cooling of a gas of solar composition²⁷. The most important point is that, although the bulk Earth does not have an exactly CI composition²⁸, refractory lithophile elements such as Ca, Sc, Ti and the rare earths are present in the silicate Earth in the same relative proportions as in the carbonaceous chondrites^{3,26,29}. Carbonaceous chondrites represent undifferentiated protoplanetary material, so the implication is that the bulk Earth contains the same relative proportions of all refractory elements as carbonaceous chondrites. This is the basis of the chondritic reference model.

As can be seen on the left-hand side of the diagram, there is a decreasing relative abundance of elements in the silicate Earth with decreasing condensation temperature or increasing volatility. This demonstrates that the Earth is depleted in volatile elements with respect to the chondritic reference, although the correlation with increasing volatility is purely qualitative. Depletions of the silicate Earth in refractory siderophile elements such as W, Mo, Re and Os are due to partial extraction of these elements into the core. The extent of extraction can be estimated by comparing silicate Earth concentrations (ratioed to CI chondrites) to those of refractory lithophile elements such as the rare-earth elements. This enables us to calculate, for each element, a core–mantle partition coefficient D_i defined as follows:

$$D_i = \frac{[i]_{\text{core}}}{[i]_{\text{silicate Earth}}}$$

where $[i]$ is the concentration of element i . Lithophile elements have D values close to zero. Table 1 shows that siderophile elements exhibit a very wide range of core–mantle partitioning behaviour, reflecting their different chemical properties and the conditions under which core segregation took place.

Core–mantle partitioning is best defined for those refractory elements that are weakly or moderately siderophile and which, like Ni and Co, are compatible in solid mantle silicates. The concentrations of these elements vary little in mantle samples, which means that D_i values based on the chondrite model have small uncertainties. Highly siderophile refractory elements such as the Pt group (Fig. 1) are slightly less well-constrained. The abundances of these elements in mantle samples are very low and, since they enter minor sulphide

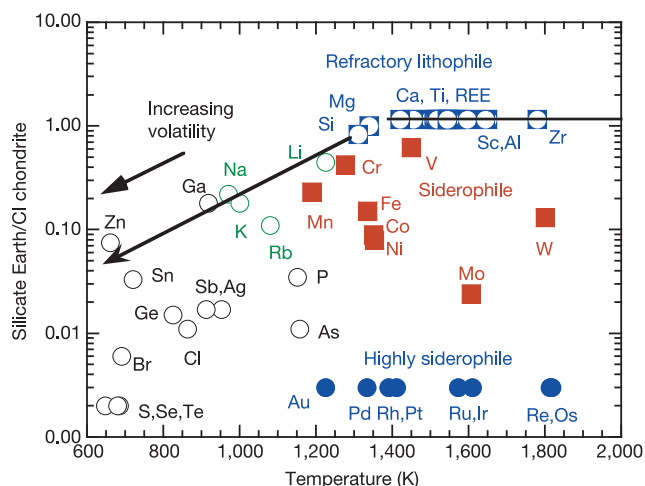


Figure 1 | Elemental abundance in the silicate Earth versus temperature of 50% condensation. Elemental abundances in the silicate Earth are ratioed to those in CI carbonaceous chondrites³ and normalized to $\frac{[\text{Mg}]_{\text{Earth}}}{[\text{Mg}]_{\text{CI}}} = 1.0$. The abundances are plotted against the temperature by which 50% of the element would have condensed from a gas of solar composition at a total pressure of 10^{-4} bar (ref. 27). We note that depletions of the silicate Earth in refractory 'Siderophile' and 'Highly siderophile' elements relative to 'Refractory lithophile' elements are due to sequestration in the core. Core contents of volatile elements that condense at low temperatures are more difficult to constrain. REE, rare-earth elements.

Table 1 | Core–mantle partition coefficients

Element	Refs 3 and 29	Ref. 26	Likely range	Low-pressure experimental D_i (refs 30 and 31)
D_{Fe}	13.66	13.65	13.65	13.65
D_{Ni}	26.5	24.4	23–27	4,900
D_{Co}	23.8	24.7	23–27	680
D_{V}	1.83	NE	1.5–2.2	0.02
D_{W}	16	NE	15–22	3
D_{Pd}	800	NE	600–1,000	7×10^5
D_{Ir}	800	NE	600–1,000	10^{11}
D_{Pt}	800	NE	600–1,000	4×10^6
D_{Nb}	NE	NE	0.2–0.8	NE
D_{Cr}	3.4	2.9	0.5–3.5*	0.2
D_{Mn}	0.29	5	0.2–2.0*	0.006
D_{Si}	0.29	0.34	0.1–0.35*	10^{-5}
D_{S}	76	NE	50–100*	50
D_{Ga}	NE	NE	0–1.5*	15
D_{P}	22	45	20–50*	30

NE, not estimated.

*Value uncertain due to volatility.

minerals, rather variable. It is generally accepted, however, that their ratios to one another in the silicate Earth are, as shown in Fig. 1, approximately chondritic.

Core concentrations of volatile siderophile elements such as Si, S and P are difficult to estimate because their bulk Earth contents are not well constrained. This means, for example that depletions in Si and Cr (Fig. 1) could be due to volatility or to sequestration in the core or both. The lack of constraint means that the uncertainties in D_i are large (Table 1).

Segregation of the reduced core from the oxidized mantle took place at high temperatures as shown above, and element partitioning between the two depended on oxygen fugacity, as can be seen from the redox reaction:



This reaction represents reduction of the element as it transfers from its normal oxidation state (n) in the silicate to oxidation state zero in the metal. Given a concentration of FeO in the mantle of about 8 weight per cent (wt%) and of Fe in the core of 85 wt% (refs 3, 26, 29), reaction (1) implies that the core separated from the mantle at an oxygen fugacity approximately $2 \log f_{\text{O}_2}$ units below Fe–FeO (iron-wüstite, IW) equilibrium.

Table 1 shows a comparison of the calculated core (metal)–mantle (silicate) partition coefficients and those obtained experimentally at low pressures (0–2 GPa), high temperatures ($\sim 1,800$ K) and oxygen fugacity corresponding to 8 wt% FeO in the mantle^{30,31}. As can be seen, the observed core–mantle partition coefficients are, for many of the siderophile elements, much (orders of magnitude) smaller than those determined experimentally. This manifests itself in the concentrations of many elements of concern (such as Ni, Co, Pd) being much greater in the mantle than would be predicted from the experimental data³², the so-called ‘excess siderophile problem’. Of equal importance are the weakly siderophile elements such as V and Cr for which the mantle concentrations are lower than predicted from the experimental data. Clearly, the mantle concentrations of these two groups of elements cannot be explained by an equilibrium with the metallic core at the fixed pressure, temperature and oxygen fugacity conditions of Table 1. We would expect, however, that early core segregation in small planetesimals and planetary embryos, including the proto-Earth, would have resulted in siderophile element abundances generally consistent with the D values given in Table 1. The Earth has apparently inherited little of the geochemistry of these earlier events.

Explanations for the ‘excess siderophile problem’

The W-isotopic anomalies of iron and silicate meteorites show that asteroidal bodies segregated metallic cores very rapidly, within about 5 Myr (ref. 2) after the beginning of the Solar System. The Earth, on the other hand, appears to have undergone more protracted core formation, the W-isotopic composition of the mantle implying single-stage separation of the core about 30 Myr (refs 2 and 19) after the beginning of the Solar System or, using an exponential accretion model³³, a mean-life of accretion and core-separation of 12 Myr. The implications of the W-isotopic data are that the asteroidal cores re-equilibrated with the silicate Earth during accretion. If re-equilibration had not occurred, the silicate Earth would have inherited the low-pressure core–mantle partitioning of the small bodies and also their large positive W-isotopic anomalies. Given that the ‘excess siderophile problem’ was not inherited, but arose from the manner of core formation on the Earth, several possible explanations have been proposed:

(1) Inefficient core formation^{34,35}. This is a disequilibrium model in which incomplete core separation left some fraction of core material mixed with the mantle. Later re-oxidation and re-mixing of this metal led to the observed elevated mantle concentrations of siderophile elements. Although this model can fit many of the elemental abundances, it cannot explain those of the weak side-

rophiles V and Cr³⁰. A further objection is the enormous amount of water, about 15 times the current mass of the hydrosphere, which would be required to re-oxidize the ‘retained’ metal⁸.

(2) Heterogeneous accretion^{8,36,37}. This model proposes that the conditions of core segregation changed from reducing at the beginning of accretion to more oxidized towards the end. During the earliest, reduced, phase, highly and moderately siderophile elements (Ni, Co and the Pt group, for example) would have been completely extracted to the core, together with some fraction of the weak siderophiles such as Si, V and Cr. As conditions became more oxidizing, core extraction of siderophile elements would have ceased progressively, beginning with the weakest siderophiles and ending with more strongly metallic elements. The final phase, after core formation had ceased, was the addition of a ‘late veneer’ of about 0.5% of chondritic material. The latter raised the concentrations of highly siderophile elements in the mantle and, because core formation had ceased, ensured their approximate chondritic proportions to one another (Fig. 1). This model solves the ‘excess siderophile problem’, but, being a disequilibrium model with multiple steps, is very difficult to test.

(3) High-pressure core formation. Extension of metal–silicate partitioning experiments to pressures above 3 GPa (refs 4, 5, 38–41) demonstrated that the partition coefficients of some siderophile elements (most notably Ni and Co; Fig. 2) change with increasing pressure such that their mantle abundances may be explained by metal–silicate equilibrium at very high pressures and temperatures (see refs 4 and 42 for example). These observations led to the ‘deep magma ocean’ hypothesis to explain the ‘excess siderophile’ problem. As the Earth grew, it is argued (see ref. 43 for example), droplets of metallic liquid descended through a 700–1,200-km-deep (28–40 GPa) magma ocean, equilibrating with the silicate liquid as they fell⁴⁴. The liquid metal ponded at the base of the magma ocean and subsequently descended in large diapirs to the growing core without further equilibration with the surrounding silicate (Fig. 3). High-pressure core formation can explain, by metal–silicate equilibrium, the partitioning of many elements between core and mantle without recourse to *ad hoc* assumptions about the oxidation state of the growing Earth. The ‘late veneer’ of chondrite-like material is, however, still required to explain the concentrations of the highly siderophile elements in the silicate Earth.

The ‘deep magma ocean’ model of core formation

As the Earth grew, the gravitational energy deposited by accreting planetesimals increased and at about 10% of its current size⁴⁵ was sufficient for significant melting to occur. After this point the

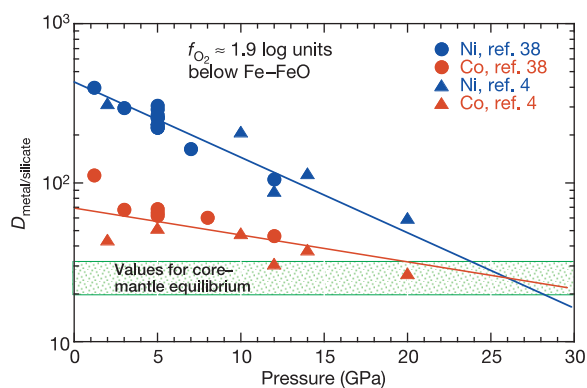


Figure 2 | The effect of pressure on Ni and Co partitioning. The liquid metal–liquid silicate partition coefficients ($D_i = \frac{[i]_{\text{metal}}}{[i]_{\text{silicate}}}$) for Ni and Co are those at 2,123–2,750 K and oxygen fugacity corresponding to current FeO content of the mantle. Data are from ref. 4 and ref. 38. Note that values approach those required for core–mantle equilibrium at the current FeO content of the mantle when pressure is ~ 28 GPa.

Earth would have periodically had an extensively molten outer layer (a magma ocean) of variable thickness. In some cases the impact energy would have been sufficient to melt both the impactor and the proto-Earth^{46,47}. The fates of pre-existing cores depended on the sizes of the Earth and impactor at the times of impact. Large impacts may have allowed the accreting core to combine directly with that of the proto-Earth. This cannot have been common, however, because of the evidence of metal–silicate re-equilibration on the Earth. Most impactors disaggregated and metallic iron sank through the molten silicate layer in droplets which, because of Rayleigh–Taylor instabilities⁴⁴, would have been about 1 cm in diameter. Droplets of this small size would have re-equilibrated with silicate melt as they fell through depths of only 60 m (ref. 44). Liquid metal and silicate therefore continued to re-equilibrate until the former either reached the core–mantle boundary (if the mantle were completely molten) or collected at a level above a solid, high-viscosity lower layer (Fig. 3). In the latter case the metal layer would cease re-equilibrating after it had reached about 5 km in thickness⁴⁴ and because of the high viscosity of the lower layer would segregate as large diapirs to the core⁴⁸.

Numerical simulations^{49,50} indicate that a magma ocean extending to the core–mantle boundary would be short-lived and that the lower mantle would crystallize in a few thousand years. A shallower, partially molten layer would crystallize much more slowly, however, and could remain as a mixture of crystals and melt for 100 Myr (refs 49, 50). Considering these results and the energetics of impact and core segregation leads to a dynamic view of the growing Earth in which the outer molten part deepened and shallowed many times after episodic impact. The pressures and temperatures recorded by core–mantle partitioning are therefore values averaged over numerous cycles of metal accumulation and segregation such as that depicted in Fig. 3.

Inspection of Fig. 3 and consideration of pressures within a

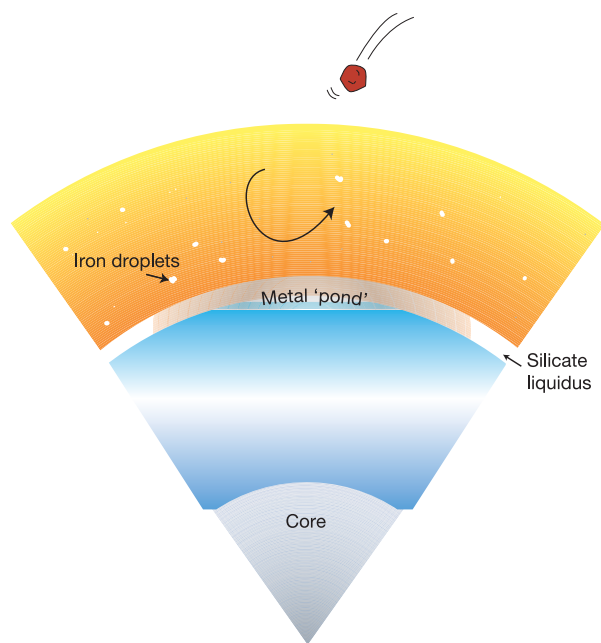


Figure 3 | The deep magma ocean model. Impacting planetesimals disaggregate and their metallic cores break up into small droplets in the liquid silicate owing to Rayleigh–Taylor instabilities. These droplets descend slowly, re-equilibrating with the silicate until they reach a region of high viscosity (solid), where they pond in a layer. The growing dense metal layer eventually becomes unstable and breaks into large blobs (diapirs), which descend rapidly to the core without further interaction with the silicate. Note that the liquidus temperature of the silicate mantle should correspond to pressure and temperature conditions at a depth above the lower solid layer and plausibly within the metal layer as indicated.

growing planet lead to several conclusions about the conditions of core segregation. First, accumulation at the base of a completely or partially molten layer implies temperatures close to or slightly below the liquidus of mantle peridotite. Second, average pressure should have increased as the planet grew. Third, because the liquidus has a positive pressure–temperature slope, the average temperature of core segregation should also have increased as the planet grew. Having accepted these principles, we can use the dependences of D_i on pressure and temperature^{6,39,40,51} to develop constraints on the core-segregation process. As shown in Fig. 2, the metal–silicate partitioning behaviour of both Ni and Co is strongly dependent on pressure, with temperature effects being subordinate^{6,52} while partitioning of V and Si, for example, depends predominantly on temperature⁶. Therefore, by simultaneous consideration of a number of elements of differing chemical behaviour, it is possible to test potential pressure–temperature paths of the accretion and core-segregation process.

Accretion and core-segregation processes

For the simplest end-member case of single-stage equilibration of mantle and core, current partitioning data for Ni and Co yield pressures close to 40 GPa while temperature sensitive elements such as V and Cr yield temperatures of about 3,750 K (ref. 6). These conditions result in a core with about 10.5 wt% Si as the dominant light element⁶ but, as illustrated in Fig. 4, correspond to temperatures about 650 K above the liquidus of the mantle. Such temperatures are physically implausible, however, because the base of the magma ocean must be close to the mantle liquidus (Fig. 3). If the FeO content of the mantle (and hence the oxygen fugacity) is fixed, then forcing core–mantle equilibration to remain on the peridotite liquidus results in an overall vanadium (V) core–mantle partition coefficient of about 0.4, a factor of four lower than that required. Because V partitioning is insensitive to pressure^{6,51}, this is purely a temperature effect—peridotite liquidus temperatures are too low to partition the requisite amount of V into the core. V partitioning into the metal can be enhanced by assuming that the silicate is a crystal–liquid mixture but the effect is small. Furthermore, crystal–liquid mixtures tend to partition too much Nb and W into the core. The

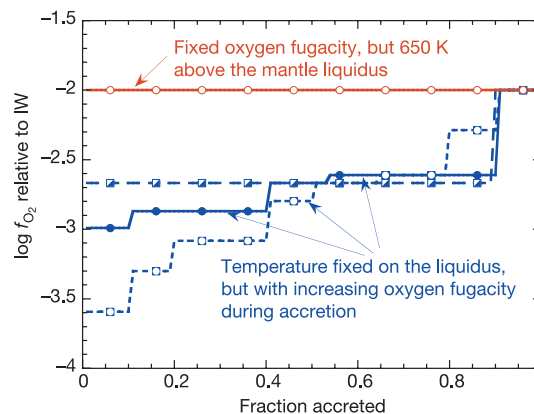


Figure 4 | Conditions yielding correct core–mantle partitioning of siderophile elements during accretion. Oxygen fugacity (relative to IW or Fe–FeO) is plotted as a function of fraction accreted, for Earth accretion models that yield the expected D values of Table 1 for Fe, Ni, Co, V, W, Nb, Cr, Mn, Ga, Si and P. Accreting metal was assumed to equilibrate with the silicate mantle at the base of the magma ocean, which deepens as the Earth grows. Note that at fixed oxygen fugacity (or FeO content of the mantle) during accretion, a temperature about 650 K above the silicate liquidus is required. Metal–silicate equilibration models fixed at the silicate liquidus temperature during accretion require pressures (from Ni and Co partitioning) corresponding to 30–40% of the depth to the core–mantle boundary in the growing planet and also that oxygen fugacity increases as the planet grows.

simplest solution⁶, as illustrated in Fig. 4, is to allow oxygen fugacity (as represented by the FeO content of the mantle) to increase as accretion and core-segregation progress.

Figure 4 illustrates a continuous process of accretion and core segregation in 1% intervals. New material added to the Earth was assumed to be mixed with the pre-existing mantle, but isolated from the proto-core. The core was segregated in 1% steps, each aliquot being in equilibrium with the entire mantle at the time of segregation at pressures constrained by Ni and Co partitioning and temperatures on the peridotite liquidus⁶. The latter were found to be consistent with magma oceans extending to 30–40% of the depth of the growing mantle. A large number of oxygen fugacity paths (of which three are illustrated) yield the partitioning of V, W, Nb, Ni, Co, Mn, Cr, Ga, P and Si, consistent with the estimates of Table 1. All require, however, that most of the Earth formed under conditions more reducing than those implied by the current FeO content of the mantle and that the Earth became more oxidized towards the end of accretion.

Additionally, although this is an area of continuing debate⁵³, these accretion models require addition of 0.5% of chondritic material after cessation of core formation to generate the chondritic ratios of the highly siderophile elements such as Pd, Ir and Pt in the silicate Earth (Fig. 1). The current 'best explanation' is therefore an amalgam of 'deep magma ocean' and 'heterogeneous accretion' hypotheses, that is, core segregation at the base of a deepening magma ocean under progressively more oxidizing conditions⁶. The important questions then are: what was the mechanism of oxidation and when did it operate? That oxidation has occurred is demonstrated by the current high oxygen fugacity of the mantle⁵⁴.

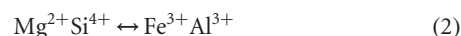
Oxidation of the Earth during and after accretion

The oxidation state of the Earth would have increased during accretion if late-arriving planetesimals had higher FeO/Fe ratios than the early ones⁸, thus increasing the oxidized iron content of the mantle and generating a concomitant increase in oxygen fugacity. This idea, although impossible to quantify, is broadly consistent with planetesimal theory, which states that during the period of oligarchic

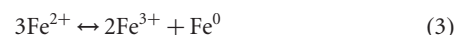
growth the proto-Earth consumed bodies in a narrow feeding zone (0.01 astronomical units, AU) but that the source of accreting bodies dramatically widened in the later stages of planetary growth¹. It is thus conceivable that late accretion added more oxidized materials from the current position of the asteroid belt¹.

Dissociation of H₂O and escape of H₂ from the Earth's atmosphere⁵⁵ may also have caused oxidation. If this is an important oxidation mechanism, however, it is difficult to see why the mantle of Mars, a more volatile-rich planet than the Earth, with a higher FeO content³⁷, appears to be more reduced than the Earth's mantle⁵⁶. Finally, a recent hypothesis^{6,7} (elaborated below) is that the Earth simply self-oxidized through crystallization of magnesium silicate perovskite.

At depths below 660 km in the present-day Earth, the stable phases are, assuming peridotitic composition, (Mg,Fe)SiO₃, magnesium silicate perovskite (79% by volume, vol.%; ref. 57), (Mg,Fe)O magnesiowüstite (16 vol.%) and CaSiO₃ perovskite (5 vol.%). An important property of magnesium silicate perovskite is that it dissolves the 4% Al₂O₃ present in peridotite by a coupled substitution with Fe³⁺ (refs 58 and 59) as follows:



It has recently been discovered that this substitution mechanism is so stable that it forces ferrous iron to disproportionate to ferric iron plus iron metal⁷:



Or, in terms of oxide components:



This means that, as perovskite began to crystallize in the extensively molten Earth, it dissolved ferric iron as FeAlO₃ component and produced Fe metal. The perovskite is stable above 23 GPa, so this process took place over a substantial part of the history of the Earth's accretion and core formation. The implications are that metal sinking through the lower mantle to the core would inevitably have dissolved some of this internally produced Fe, resulting in a perovskitic layer which was relatively oxidized. Given the gravitational instability of any metal layer (Fig. 3), accretional energies and heat-loss, the depth of the magma ocean must have fluctuated continuously, thereby generating fronts of dissolution and precipitation at the lower boundary, as depicted in Fig. 5.

Perovskite dissolution and re-precipitation acted as a ferric or oxygen 'pump', releasing ferric iron to the magma ocean during dissolution and producing more during precipitation. As the Earth grew, ferric iron released to the magma ocean would have been consumed by reduced species such as CH₄. Droplets of Fe metal falling through the magma ocean would also have reacted with ferric iron, driving reaction (3) to the left and producing more Fe²⁺ to dissolve in the silicate melt. Hence the oxygen fugacity and the content of oxidized iron in the mantle (magma ocean) increased naturally as a consequence of perovskite crystallization and dissolution. In the very final stages of earth accretion this mechanism may have caused sufficient oxidation to halt metal segregation, setting the stage for the 'late veneer' of chondritic or similar material to add the highly siderophile elements exclusively to the mantle. The importance of this self-oxidation mechanism is that it is experimentally observed to take place when magnesium silicate perovskite crystallizes⁷.

The self-oxidation mechanism removes the need to make *ad hoc* assumptions about the compositions of the materials added to the Earth during accretion. The oxidation-state changes required by partitioning data are a simple consequence of the size of the Earth. Any planet in which magnesium silicate perovskite is an important crystallizing phase should undergo the same process. It could explain why the Earth's mantle is more oxidized than that of Mars⁵⁶, a planet in which perovskite is either absent or only stable very close to the

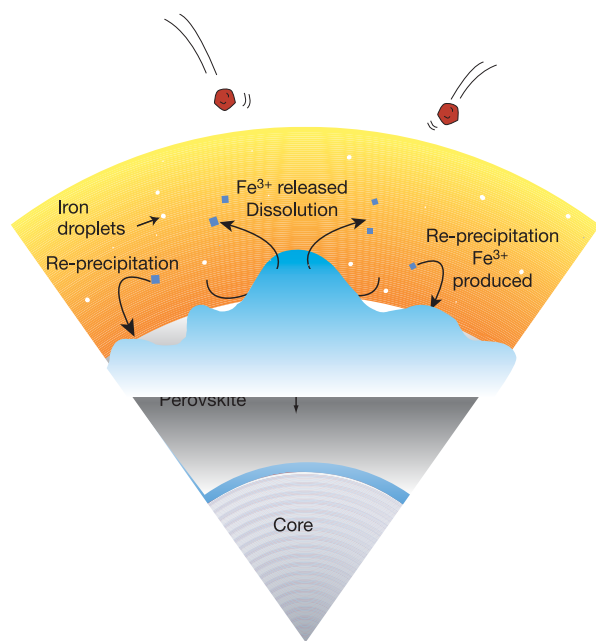


Figure 5 | Sketch of a possible mechanism by which the mantle may have self-oxidized via perovskite crystallization. Lower-mantle perovskite (Pv) generates Fe³⁺ and Fe⁰ (metal) from disproportionation of Fe²⁺. Fe metal is swept to the core by descending metal diapirs while the vigour of accretion and core segregation causes dissolution and recrystallization of perovskite with release of Fe³⁺ to the overlying magma ocean.

core–mantle boundary. Because of its small size, Mars cannot have undergone the same period of extensive perovskite crystallization during accretion as did the Earth.

Consequences of 'self-oxidation' and the age of the core

The timing of core segregation can, in principle, be determined from a radioactive decay system in which either the parent or the daughter is siderophile, so that core formation established the current parent/daughter ratio in the silicate Earth. The systems ^{182}Hf – ^{182}W and $^{238,235}\text{U}$ – $^{206,207}\text{Pb}$ are potential core chronometers. The daughters, W and Pb, are siderophile and their current isotopic constitutions yield model-dependent 'ages' of core formation. Ages are obtained by assuming, for example, instantaneous core segregation at a fixed time or, more realistically, continuous core segregation as the Earth grew.

If it is assumed that all accreting material mixed isotopically with the silicate Earth, that accretion rate decreased exponentially with increasingly larger impacts, as predicted from the planetesimal theory and that the core segregated continuously with fixed parent and daughter partitioning, a time constant for core formation can be obtained (Fig. 6):

$$F_t = 1 - e^{-\lambda \Delta t} \quad (5)$$

where F_t is the cumulative fractional mass of the Earth at time t relative to the present day, λ is the time constant for core formation ($= 1/\tau$, where τ = the mean life) and $\Delta t = t_0 - t$ (where t_0 is the age of the Solar System). On this basis, both Hf–W and U–Pb timescales should be the same, but as can be seen from Fig. 6, the mean-life of core formation is about 12 Myr using the Hf–W system and 28 Myr using U–Pb. Changing the model (for example, to single-stage equilibration) alters both timescales, but the discrepancy between them remains. The U–Pb timescale is always more protracted than that of Hf–W.

The only way in which both Hf–W and U–Pb timescales could be

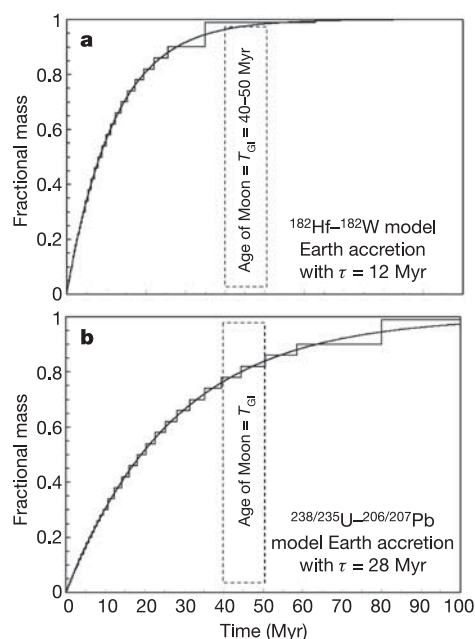


Figure 6 | Two estimates of the timing of accretion and core formation on the Earth. Both estimates assume an exponential growth model (equation (5) with progressively increasing masses of impactors). **a**, The result that satisfies the W isotopic composition of the bulk silicate Earth⁹. **b**, A result that satisfies recent estimates of the Pb isotopic composition of the bulk silicate Earth⁹. The Hf–W age of the Moon corresponding to the Moon-forming giant impact (T_{GI}) appears to lie in the range 40 to 50 Myr (ref. 61). As discussed in the text, the W result dates the principal phase of core formation, while the Pb result plausibly dates a late-stage segregation of sulphide after the giant impact.

correct is if Pb entered the core at a later stage than W, after the formation of the Moon⁹. The giant impact believed to have formed the Moon^{46,60} defined the last major growth stage of the Earth, adding about 10% of its mass. The timing of the impact is given by the Hf–W age of the Moon which appears to lie between 40 and 50 Myr (Fig. 6)⁶¹. Under the reducing conditions of early accretion (Fig. 4) W would be a siderophile element while Pb partitioning into the metal phase would be weak^{9,62}. Conditions appear to have become more oxidizing during accretion, however, so that, as found in modern MORB⁶³, an Fe–Ni sulphide became the only stable 'metallic' phase (Fig. 7). Figure 7 shows the effect when ferric iron produced from perovskite crystallization is recycled back into a melt of peridotitic composition under upper mantle conditions. Oxygen fugacity increases by about three log units and sulphur-rich liquids are the only metallic phases which can coexist with the putative magma ocean.

The giant Moon-forming impact should have completely melted the Earth⁶⁰, dramatically increasing the solubility of sulphur and chalcophile elements in the liquid silicate⁹. As the Earth cooled, however, an FeS-rich liquid would have 'rained out' of the liquid mantle, extracting chalcophile elements. Importantly, Pb is chalcophile, partitioning strongly into iron sulphide liquid⁶⁴. Segregation of <1% of late sulphide liquid, the 'Hadean Matte'⁸, would therefore have fractionated U from Pb, re-setting the U–Pb age without affecting the Hf–W age because the ^{182}Hf parent was already extinct⁹. Thus the U–Pb age of the Earth (65–85 Myr after the beginning of the Solar System) plausibly represents the very last stages of core segregation at the end of accretion, while the W age represents the major phase of core segregation before the Moon-forming impact.

The 'light' element in the core

Although the Earth's core is, in Birch's⁶⁵ words, an 'uncertain mixture of all the elements', we are able through geochemical and geophysical arguments to constrain its composition quite well. The chondritic reference model (Fig. 1), when combined with the compositions of carbonaceous chondrites⁶⁶ and the primitive mantle^{3,26}, leads to a core which is about 85% Fe and 5% Ni by weight. All of the other refractory elements discussed here should be present at very low concentration in the core, with the exceptions of Cr ($\leq 0.9\%$) and Co

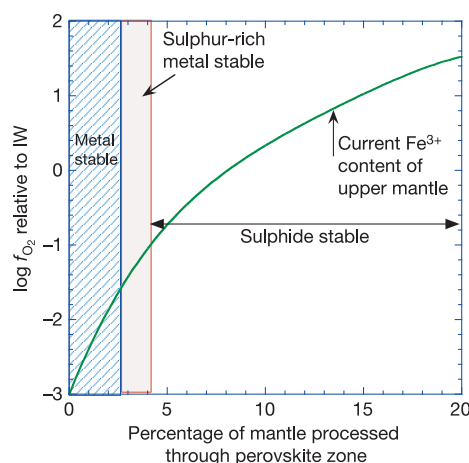


Figure 7 | Calculated effect of perovskite crystallization on the Fe^{3+} content and hence oxygen fugacity of a magma ocean of peridotite composition. The calculation began by assuming peridotitic mantle of current composition, containing 8% FeO. Perovskite crystallized from a melt of this composition contains about 3.5% 'FeO' but rather than being entirely ferrous, half of the iron is actually ferric^{7,77}. The Fe^{3+} content of the molten mantle increases as perovskite crystallizes, then dissolves, releasing Fe^{3+} , and finally recrystallizes at the base of the magma ocean. This 'processing' of the mantle was converted from Fe^{3+} content to oxygen fugacity using experimental data on silicate melts⁷⁸. Oxygen fugacities are expressed relative to the iron–wüstite (Fe–FeO) buffer.

(0.25%). This means that the core contains about 8 wt% of non-refractory elements and this 8% also corresponds reasonably well to the density deficit of the core compared to pure Fe–Ni alloy under core conditions^{45,67}.

Because only elements of low atomic number can significantly lower the density of an Fe-rich core, the ‘missing’ 8% of core mass is generally referred to as the ‘light’ element. The most likely candidates are low-atomic-number elements that are soluble in Fe and that are cosmically abundant. This leads us to H, O, C, S, P and Si as possible ‘light’ substituents in the core⁶⁸. Of these, the clearest case can be made for sulphur, which is a strongly siderophile element. Sulphur is moderately volatile (Fig. 1), but is almost two orders of magnitude more depleted in the silicate Earth than an element of similar volatility, zinc (Zn) (Fig. 1). Making the reasonable assumption that this depletion of S relative to Zn in the silicate Earth is due to core formation gives us a core concentration of 1.9 wt% S (refs 26, 29, 69). The same procedure applied to phosphorus leads to about 0.2% of this element in the core²⁹. Hydrogen and carbon are two cosmically abundant siderophile elements which are also highly volatile and hence strongly depleted in the Earth. A rough estimate of their core concentrations based on an extrapolation of the volatility trend of Fig. 1 leads to 0.1% and 0.2%, respectively²⁹. This leaves the most controversial of the candidates, oxygen and silicon to make up the balance.

The depletion of Si in the silicate Earth relative to the chondritic reference (Fig. 1) is due to an unknown combination of volatility (incomplete accretion) and dissolution in the core. If dissolution in the core were the principal reason then the latter would contain 6–7 wt% Si (refs 26 and 29) and Si would be the major ‘light’ component. Low-pressure core segregation cannot achieve such high Si concentrations without conditions being much too reducing for the current FeO content of the mantle⁷⁰. Si solubility in liquid Fe increases with both pressure and temperature^{6,71}, however, such that significant amounts would dissolve at pressures above 20 GPa, even at the current FeO content of the mantle. If we assume that, as discussed earlier (Fig. 4), most of the core formed under slightly more reducing conditions, then the continuous growth and segregation models (Fig. 4) lead to 4–5 wt% Si in the core, which implies that, without making any explicit assumptions about volatility, Si is an important constituent. The oxygen content of the core is more difficult to constrain because the available experimental data conflict with one another. At low pressures and temperatures, oxygen solubility in liquid iron is negligible⁷². All studies show that the effect of temperature is to increase solubility dramatically, however^{72,73}, so that at low pressures and 2,800 K iron liquid in equilibrium with liquid mantle would contain about 4 wt% oxygen⁷³.

The main controversy concerns the effect of pressure. Ohtani *et al.*⁷⁴ suggested that the effect of pressure is to increase oxygen solubility in Fe, whereas most recent studies^{73,75} have found the opposite effect. If we take the combined effects of temperature and pressure as found by Rubie *et al.*⁷³, and apply them to metal–silicate equilibration on the peridotite liquidus, we find oxygen concentrations of up to about 1 wt% in the metal. Lowering of oxygen fugacity during most of accretion decreases the oxygen content of the metal. Thus, if we accept the data indicating that pressure reduces the solubility of oxygen in metallic iron, then the oxygen content of the core should be <1%. This is an area of vigorous current debate, however, and a recent study in the diamond anvil cell⁷⁶ found the opposite effect of pressure on oxygen solubility. Furthermore, at approximately the appropriate oxygen fugacity for core–mantle equilibrium, Takafuji *et al.*⁷⁶ found oxygen concentrations in the metal to be about 1.5 times the Si concentrations. Clearly Si is likely to be present at moderate concentrations in the core and oxygen may also be a significant substituent.

Conclusions

Chemical and isotopic measurements on the silicate part of the Earth combined with astronomical observations of young stars and

physical models of planetary development lead to a coherent view of the early history of the Earth. The energetics of accretion⁴⁵ imply continuous core segregation from a partially molten mantle under conditions that generated the core–mantle partition coefficients of Table 1. The siderophile elements Ni and Co have strongly pressure-dependent partitioning and point towards a mean pressure of core segregation greater than 30 GPa. Partitioning of more highly charged weak siderophiles such as V, Cr, Si and Nb provide better constraints on mean temperature and oxygen fugacity of core formation. Thus, at pressures constrained principally by Ni and Co, these elements point to temperatures 650 K above the peridotite liquidus (Fig. 4). To reduce the required temperature to the liquidus, consistent with the deep magma ocean model (Fig. 3), some part of core formation must have taken place under conditions more reducing than those implied by the current FeO content of the mantle⁶ (Fig. 4). Thus, current best estimates of metal–silicate partitioning are consistent with an important component of the ‘heterogeneous accretion’ model, that oxygen fugacity increased during accretion. The constraints provided by the partitioning data would, however, be stronger if there were more experimental data at pressures above 15 GPa, particularly for the weakly siderophile elements.

Following the demonstration that lower-mantle perovskite forces disproportionation of Fe²⁺ to Fe³⁺ plus iron⁷, it has become clear that the Earth’s mantle probably ‘self-oxidized’ during core formation⁶. This means that the increase in oxygen fugacity implied by the metal–silicate partitioning data could be explained as a natural consequence of the Earth having reached a size large enough to stabilize substantial amounts of perovskite in the lower mantle (Fig. 5). The same mechanism could have operated on Venus, to a limited extent on Mars and not at all on the Moon, or any of the asteroids.

Self-oxidation during accretion has also been invoked as the reason for the difference between Hf–W and U–Pb ages of the core⁹. In this hypothesis (Fig. 6), W was partitioned into the core during the early ‘reduced’ phase of metal segregation while Pb largely remained behind in the mantle. Progressive oxidation means, it is argued⁹, that the compositions of liquid metals extracted during the last stages of core formation were sulphur-rich and Pb should partition strongly into such metals. Thus, the Hf–W age of the Earth could reasonably refer to the earliest phases of core formation and the U–Pb age to the last. To test this hypothesis properly it will be necessary to obtain experimental data on the partitioning of Pb between silicates and metals as a function of oxygen fugacity, pressure, temperature and sulphur content.

The density of the liquid outer core is about 8% lower than pure Fe alloy under core conditions, implying the presence of low-atomic-number alloying elements. Cosmochemical arguments suggest that S, H and C are present at concentrations of 1.9, 0.1 and 0.2 wt% respectively. Si is a cosmochemically abundant element which dissolves readily in iron under reducing conditions and at high pressures and temperatures. The accretion model discussed earlier, in which increases in oxygen fugacity accompany core segregation, leads (given a mantle Si content of 21 wt%; refs 3, 26) to 4–5% Si in the core. The remaining potentially important component, oxygen, is more difficult to constrain because the available experimental data conflict with one another. Estimates based on some recently published data⁷³ lead, under the conditions of core segregation implied by siderophile element partitioning, to <1 wt% oxygen in the core. Other recent data⁷⁶ imply that the core should be richer in oxygen than in silicon. Clearly, more experimental data are needed to resolve the question of the oxygen content of the core.

1. Chambers, J. E. Planetary accretion in the inner solar system. *Earth Planet. Sci. Lett.* **224**, 241–252 (2004).
2. Kleine, T., Munker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf–W chronometry. *Nature* **418**, 952–955 (2002).
3. McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120**, 223–253 (1995).

4. Li, J. & Agee, C. B. Geochemistry of mantle-core differentiation at high pressure. *Nature* **381**, 686–689 (1996).
5. Righter, K. & Drake, M. J. Metal/silicate equilibrium in the early Earth—New constraints from the volatile moderately siderophile elements Ga, Cu, P, and Sn. *Geochim. Cosmochim. Acta* **64**, 3581–3597 (2000).
6. Wade, J. & Wood, B. J. Core formation and the oxidation state of the Earth. *Earth Planet. Sci. Lett.* **236**, 78–95 (2005).
7. Frost, D. J. *et al.* Experimental evidence for the existence of iron-rich metal in the Earth's lower mantle. *Nature* **428**, 409–412 (2004).
8. O'Neill, H. S. C. The origin of the Moon and the early history of the Earth—A chemical model. 2. The Earth. *Geochim. Cosmochim. Acta* **55**, 1159–1172 (1991).
9. Wood, B. J. & Halliday, A. N. Cooling of the earth and core formation after the giant impact. *Nature* **437**, 1345–1348 (2005).
10. Youdin, A. N. & Shu, F. H. Planetsimal formation by gravitational instability. *Astrophys. J.* **580**, 494–505 (2002).
11. Poppe, T., Blum, J. & Henning, T. Analogous experiments on the stickiness of micron sized planetary dust. *Astrophys. J.* **533**, 454–471 (2000).
12. Weidenschilling, S. J. Aerodynamics of solid bodies in the solar nebula. *Mon. Not. R. Astron. Soc.* **180**, 57–70 (1977).
13. Benz, W. & Asphaug, E. Catastrophic disruptions revisited. *Icarus* **142**, 5–20 (1999).
14. Leinhardt, Z. M., Richardson, D. C. & Quinn, T. Direct N-body simulations of rubble-pile collisions. *Icarus* **146**, 133–151 (2000).
15. Wetherill, G. W. & Stewart, G. R. Formation of planetary embryos: effects of fragmentation, low relative velocity, and independent variation of eccentricity and inclination. *Icarus* **106**, 190–209 (1993).
16. Kokubo, E. & Ida, S. Oligarchic growth of protoplanets. *Icarus* **131**, 171–178 (1998).
17. Weidenschilling, S. J., Spaute, D., Davis, D. R., Marzari, F. & Ohtsuki, K. Accretional evolution of a planetesimal swarm. 2. The terrestrial zone. *Icarus* **128**, 429–455 (1997).
18. Taylor, S. R. & Norman, S. R. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 29–44 (Oxford Univ. Press, New York, 1990).
19. Yin, Q. Z. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
20. Stevenson, D. J. in *Origin of the Earth* (eds Newsom, H. E. & Drake, J. H.) 231–249 (Oxford Univ. Press, New York, 1990).
21. Yoshino, T., Walter, M. J. & Katsura, T. Core formation in planetesimals triggered by permeable flow. *Nature* **422**, 154–157 (2003).
22. Ghosh, A. & McSweeney, H. Y. A thermal model for the differentiation of Asteroid 4 Vesta. *Icarus* **134**, 187–206 (1998).
23. Shukolyukov, A. & Lugmair, G. W. Live iron-60 in the early solar system. *Science* **259**, 1138–1142 (1993).
24. Kunihiro, T., Rubin, A. E., McKeegan, K. D. & Wasson, J. T. Initial $^{26}\text{Al}/^{27}\text{Al}$ in carbonaceous-chondrite chondrules: Too little ^{26}Al to melt asteroids. *Geochim. Cosmochim. Acta* **68**, 2947–2957 (2004).
25. Chambers, J. E. & Wetherill, G. W. Making the terrestrial planets: N-body integrations of planetary embryos in three dimensions. *Icarus* **136**, 304–327 (1998).
26. Allegre, C. J., Poirier, J.-P., Humler, E. & Hofmann, A. W. The chemical composition of the Earth. *Earth Planet. Sci. Lett.* **134**, 515–526 (1995).
27. Wasson, J. T. *Meteorites: Their Record of Early Solar-System History* (W. H. Freeman, New York, 1985).
28. Drake, M. J. & Righter, K. Determining the composition of the Earth. *Nature* **416**, 33–39 (2002).
29. McDonough, W. F. in *The Mantle and Core* (ed. Carlson, R. W.) 547–568 (Elsevier-Pergamon, Oxford, 2003).
30. Newsom, H. E. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 273–288 (Oxford Univ. Press, New York, 1990).
31. O'Neill, H. S. C. & Palme, H. in *The Earth's Mantle—Composition, Structure and Evolution* (ed. Jackson, I.) 3–127 (Cambridge Univ. Press, Cambridge, UK, 1998).
32. Ringwood, A. E. Chemical evolution of the terrestrial planets. *Geochim. Cosmochim. Acta* **30**, 41–104 (1966).
33. Harper, C. L. Jr & Jacobsen, S. B. Evidence for ^{182}Hf in the early Solar System and constraints on the timescale of terrestrial accretion and core formation. *Geochim. Cosmochim. Acta* **60**, 1131–1153 (1996).
34. Arculus, R. J. & Delano, J. W. Siderophile element abundances in the upper mantle: Evidence for a sulfide signature and equilibrium with the core. *Geochim. Cosmochim. Acta* **45**, 1331–1343 (1981).
35. Jones, J. H. & Drake, M. J. Geochemical constraints on core formation in the Earth. *Nature* **322**, 221–228 (1986).
36. Chou, C. L. Fractionation of siderophile elements in the Earth's upper mantle. *Proc. Ninth Lunar Sci. Conf.*, 219–230 (1978).
37. Wänke, H. Constitution of the terrestrial planets. *Phil. Trans. R. Soc. Lond. A* **303**, 287–302 (1981).
38. Thibault, Y. & Walter, M. J. The influence of pressure and temperature on the metal-silicate partition coefficients of nickel and cobalt in a model C1 chondrite and implications for metal segregation in a deep magma ocean. *Geochim. Cosmochim. Acta* **59**, 991–1002 (1995).
39. Righter, K., Drake, M. J. & Yaxley, G. Prediction of siderophile element metal-silicate partition coefficients to 20 GPa and 2800 degrees C: The effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt compositions. *Phys. Earth Planet. Inter.* **100**, 115–134 (1997).
40. Li, J. & Agee, C. B. The effect of pressure, temperature, oxygen fugacity and composition on partitioning of nickel and cobalt between liquid Fe-Ni-S alloy and liquid silicate: Implications for the Earth's core formation. *Geochim. Cosmochim. Acta* **65**, 1821–1832 (2001).
41. Gessmann, C. K. & Rubie, D. C. The effect of temperature on the partitioning of nickel, cobalt, manganese, chromium, and vanadium at 9 GPa and constraints on formation of the Earth's core. *Geochim. Cosmochim. Acta* **62**, 867–882 (1998).
42. Righter, K. & Drake, M. J. Metal-silicate equilibrium in a homogeneously accreting earth: New results for Re. *Earth Planet. Sci. Lett.* **146**, 541–553 (1997).
43. Righter, K. & Drake, M. J. Effect of water on metal-silicate partitioning of siderophile elements: a high pressure and temperature terrestrial magma ocean and core formation. *Earth Planet. Sci. Lett.* **171**, 383–399 (1999).
44. Rubie, D. C., Melosh, H. J., Reid, J. E., Liebske, C. & Righter, K. Mechanisms of metal-silicate equilibration in the terrestrial magma ocean. *Earth Planet. Sci. Lett.* **205**, 239–255 (2003).
45. Stevenson, D. J. Models of the Earth's Core. *Science* **214**, 611–619 (1981).
46. Cameron, A. G. W. & Benz, W. Origin of the Moon and the single impact hypothesis IV. *Icarus* **92**, 204–216 (1991).
47. Canup, R. M. & Asphaug, E. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412**, 708–712 (2001).
48. Karato, S.-i. & Rama Murthy, V. Core formation and chemical equilibrium in the Earth-I. Physical considerations. *Phys. Earth Planet. Inter.* **100**, 61–79 (1997).
49. Abe, Y. Thermal and chemical evolution of the terrestrial magma ocean. *Phys. Earth Planet. Inter.* **100**, 27–39 (1997).
50. Solomatov, V. S. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 323–360 (Univ. Arizona Press, Tucson, 2000).
51. Chabot, N. L. & Agee, C. B. Core formation in the Earth and Moon: new experimental constraints from V, Cr, and Mn. *Geochim. Cosmochim. Acta* **67**, 2077–2091 (2003).
52. Chabot, N. L., Draper, D. S. & Agee, C. B. Conditions of core formation in the Earth: Constraints from nickel and cobalt partitioning. *Geochim. Cosmochim. Acta* **69**, 2141–2151 (2005).
53. Cottrell, E. & Walker, D. Constraints on core formation from Pt partitioning in mafic silicate liquids at high temperatures. *Geochim. Cosmochim. Acta* (in the press).
54. Wood, B. J., Bryndzia, L. T. & Johnson, K. E. Mantle oxidation state and its relationship to tectonic environment and fluid speciation. *Science* **248**, 337–345 (1990).
55. Hunten, D. M. Atmospheric evolution of the terrestrial planets. *Science* **259**, 915–920 (1993).
56. Herd, C. D. K., Borg, L. E., Jones, J. H. & Papike, J. J. Oxygen fugacity and geochemical variations in the martian basalts: Implications for martian basalt petrogenesis and the oxidation state of the upper mantle of Mars. *Geochim. Cosmochim. Acta* **66**, 2025–2036 (2002).
57. Wood, B. J. Phase transformations and partitioning relations in peridotite under lower mantle conditions. *Earth Planet. Sci. Lett.* **174**, 341–354 (2000).
58. McCammon, C. Perovskite as a possible sink for ferric iron in the lower mantle. *Nature* **387**, 694–696 (1997).
59. Wood, B. J. & Rubie, D. C. The effect of alumina on phase transformations at the 660-kilometer discontinuity from Fe-Mg partitioning experiments. *Science* **273**, 1522–1524 (1996).
60. Canup, R. M. Simulations of a late lunar-forming impact. *Icarus* **168**, 433–456 (2004).
61. Kleine, T., Palme, H., Mezger, K. & Halliday, A. N. Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310**, 1671–1674 (2005).
62. Ohtani, E., Yurimoto, H. & Seto, S. Element partitioning between metallic liquid, silicate liquid, and lower-mantle minerals: implications for core formation of the Earth. *Phys. Earth Planet. Inter.* **100**, 97–114 (1997).
63. Francis, R. D. Sulfide globules in mid-ocean ridge basalts (MORB), and the effect of oxygen abundance in Fe-S-O liquids on the ability of those liquids to partition metals from MORB and komatiite magmas. *Chem. Geol.* **85**, 199–213 (1990).
64. Jones, J. H., Hart, S. R. & Benjamin, T. M. Experimental partitioning near the Fe-FeS eutectic, with an emphasis on elements important to iron meteorite chronologies (Pb, Ag, Pd, and Ti). *Geochim. Cosmochim. Acta* **57**, 453–460 (1993).
65. Birch, F. Elasticity and constitution of the Earth's interior. *J. Geophys. Res.* **57**, 227–286 (1952).
66. Newsom, H. E. in *Global Earth Physics: a Handbook of Physical Constants* (ed. Ahrens, T. J.) 159–189 (American Geophysical Union, Washington DC, 1995).
67. Anderson, O. L. & Isaak, D. G. Another look at the core density deficit of Earth's outer core. *Phys. Earth Planet. Inter.* **131**, 19–27 (2002).
68. Poirier, J. P. Light elements in the Earth's outer core—a critical-review. *Phys. Earth Planet. Inter.* **85**, 319–337 (1994).
69. Dreibus, G. & Palme, H. Cosmochemical constraints on the sulfur content in the Earth's core. *Geochim. Cosmochim. Acta* **60**, 1125–1130 (1996).
70. Kilburn, M. R. & Wood, B. J. Metal-silicate partitioning and the incompatibility of S and Si during core formation. *Earth Planet. Sci. Lett.* **152**, 139–148 (1997).

71. Gessmann, C. K., Wood, B. J., Rubie, D. C. & Kilburn, M. R. Solubility of silicon in liquid metal at high pressure: implications for the composition of the Earth's core. *Earth Planet. Sci. Lett.* **184**, 367–376 (2001).
72. Ohtani, E. & Ringwood, A. E. Composition of the core. I. Solubility of oxygen in molten iron at high temperatures. *Earth Planet. Sci. Lett.* **71**, 85–93 (1984).
73. Rubie, D. C., Gessmann, C. K. & Frost, D. J. Partitioning of oxygen during core formation on the Earth and Mars. *Nature* **429**, 58–61 (2004).
74. Ohtani, E., Ringwood, A. E. & Hibberson, W. Composition of the core. II. Effect of high pressure on solubility of FeO in molten iron. *Earth Planet. Sci. Lett.* **71**, 94–103 (1984).
75. O'Neill, H. S., Canil, D. & Rubie, D. C. Oxide-metal equilibria to 2500 degrees C and 25 GPa: Implications for core formation and the light component in the Earth's core. *J. Geophys. Res.* **103**, 12239–12260 (1998).
76. Takafuji, N., Hirose, K., Mitome, N. & Bando, Y. Solubilities of O and Si in liquid iron in equilibrium with (Mg,Fe)SiO₃ perovskite and the light elements in the core. *Geophys. Res. Lett.* **32**, L06313 (2005).
77. Tronnes, R. G. & Frost, D. J. Peridotite melting and mineral-melt partitioning of major and minor elements at 22–24.5 GPa. *Earth Planet. Sci. Lett.* **197**, 117–131 (2002).
78. Kilinc, A., Carmichael, I. S. E., Rivers, M. L. & Sack, R. O. The ferric-ferrous ratio of natural silicate liquids equilibrated in air. *Contrib. Mineral. Petrol.* **83**, 136–140 (1983).

Acknowledgements The constructive reviews of C. Agee and K. Righter are acknowledged with thanks. B.J.W. acknowledges the award of an ARC Federation Fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.J.W. (bwood@els.mq.edu.au).

ARTICLES

A common mass scaling for satellite systems of gaseous planets

Robin M. Canup¹ & William R. Ward¹

The Solar System's outer planets that contain hydrogen gas all host systems of multiple moons, which notably each contain a similar fraction of their respective planet's mass ($\sim 10^{-4}$). This mass fraction is two to three orders of magnitude smaller than that of the largest satellites of the solid planets (such as the Earth's Moon), and its common value for gas planets has been puzzling. Here we model satellite growth and loss as a forming giant planet accumulates gas and rock-ice solids from solar orbit. We find that the mass fraction of its satellite system is regulated to $\sim 10^{-4}$ by a balance of two competing processes: the supply of inflowing material to the satellites, and satellite loss through orbital decay driven by the gas. We show that the overall properties of the satellite systems of Jupiter, Saturn and Uranus arise naturally, and suggest that similar processes could limit the largest moons of extrasolar Jupiter-mass planets to Moon-to-Mars size.

Orbiting Jupiter, Saturn and Uranus are the so-called regular satellites—which orbit approximately within the planet's equatorial plane, in the same sense as the planet's rotation, and at orbital distances of up to tens of planetary radii—and much smaller irregular satellites, with more distant, inclined and elliptical orbits. Whereas irregular satellites are thought to be captured objects¹, regular satellites are believed to have arisen as a by-product of the planet's formation, within a circumplanetary disk of gas and solids^{2–4}. Outer planet satellite origin is important both for the history of these objects (including volcanic Io, Europa with its believed subsurface ocean, and organic-rich Titan) and because their existence holds clues to giant planet origins.

The number and individual masses of large satellites differ from planet to planet (Table 1). However, the systems share a distinctive trait: they each contain a similar total mass, M_T , compared to that of their central planet, M_P , with (M_T/M_P) varying only from 1.1×10^{-4} to 2.5×10^{-4} . This similarity is remarkable, and its cause unknown. The bulk compositions of the gas planets differ, and as such their satellite system masses are not a common fraction of each planet's solid (rock+ice) or gaseous (H+He) components. Additionally, $(M_T/M_P) \approx 10^{-4}$ is orders of magnitude smaller than the mass ratios of the large satellites of solid planets, with the Moon and Pluto's moon, Charon, having $(M_T/M_P) = 0.012$ and ~ 0.1 , respectively. Neither Jupiter nor Saturn, for instance, has such a proportionally massive Earth-sized satellite.

Gas planet satellite formation models have historically used as their initial condition a circumplanetary disk containing a total mass in gas and solids necessary to produce the satellites we see. Such an initial condition is, by definition, *ad hoc*, given both the rapid evolution of circumplanetary material compared to planetary formation timescales and dynamical processes that can cause much of the material supplied to such a disk to be eventually collected by the planet. Thus a planet's current satellites may provide little constraint on the total mass that was processed through its disk.

Here we consider a model in which satellites grow within an actively supplied circumplanetary disk, sustained by a time-dependent inflow of gas and solids from heliocentric orbit during the end stages of the planet's formation. We use both numerical

simulation and analytical estimates to describe the supply of material to the disk, satellite growth, and satellite orbital decay and loss caused by satellite interactions with the gas. We find that a common satellite system mass fraction of $\sim 10^{-4}$ results, independent of the total mass processed through the disk and only weakly dependent on the least certain model parameters. We also for the first time generate satellite systems with properties generally consistent with those of Jupiter, Saturn and Uranus using direct simulation.

Slow-inflow disk model

During a jovian planet's early growth, its massive gaseous atmosphere is distended to $\sim 10^2$ to 10^3 times the current sizes of Jupiter or Saturn³. The final stages of its growth involve both a slowing in the rate of gas inflow to the planet (for example, as the local supply of material is depleted or the circumsolar gas nebula begins to dissipate⁵), and the planet's gravitational contraction, which is accompanied by the formation of a circumplanetary disk^{6–8}. We model satellite growth within such a disk⁹, supplied by an inflow of gas and solids to circumplanetary orbit across a region extending roughly from the planet's surface to an outer distance of a few tens of planetary radii (Fig. 1). The disk gas diffuses viscously, causing it to radially spread. For a gas spreading time short compared to the characteristic time over which the inflow changes, the amount of gas in the disk can be estimated as a quasi-steady state between the inflow supply and removal as gas spreads inwards onto the planet and outwards to the disk's outer edge. Solids smaller than metre-scale are bound aerodynamically to the gas, and are delivered with it from heliocentric orbit to the disk. Once in circumplanetary orbit, rapid mutual collisions allow particles to grow large enough to decouple from the gas before they can be removed by aerodynamic drag⁹. Continued satellite growth then occurs with an overall rate controlled by the rate of solid inflow⁹.

Satellite growth and loss

A satellite's orbit is affected by torques resulting from its gravitational interactions with the gas. The satellite's gravity induces spiral density waves in the gas disk, and the interaction of these waves with the satellite yields a net negative torque that acts as a drag on the satellite's

¹Space Studies Department, Southwest Research Institute, Boulder, Colorado 80302, USA.

Table 1 | Regular satellite systems

Planet	Per cent high-Z mass	M_T/M_P	Number with $m_S/M_P > 10^{-8}$	Number with $m_S/M_P > 10^{-5}$	$a_{\max}/R_P$ of most distant with $m_S/M_P > 10^{-8}$	$<C_{lg}> (R_H)$
Jupiter	~3–12% ⁴¹	2.1×10^{-4}	4	4	26	15.1
Saturn	~12–30% ⁴¹	2.5×10^{-4}	7	1	59	18.4
Uranus	~75–90% ³	1.1×10^{-4}	5	4	23	14.5

Shown (left to right columns) are the planet, the planet's estimated mass fraction of high-Z elements (rock + ice), total satellite system mass (M_T) scaled to the planet mass (M_P), number of satellites having individual masses $m_S > 10^{-8}M_P$, number of large satellites having $m_S/M_P > 10^{-5}$, semi-major axis in planetary radii ($a_{\max}/R_P$) of the most distant satellite having $m_S/M_P > 10^{-8}$, and the average large satellite orbital spacing in mutual Hill radii, $<C_{lg}>$, with $<C_{lg}>$ for Saturn calculated for the Rhea-to-Titan spacing. The mutual Hill radius of two satellites of masses m_1 and m_2 and orbital radii a_1 and a_2 is $R_H = 0.5(a_1 + a_2)[(m_1 + m_2)/(3M_P)]^{1/3}$, with $\Delta a = (a_2 - a_1) = CR_H$. Jupiter's four similarly sized satellites each contain within a factor of about two of the Moon's mass, with Io and Europa (having $a/R_P = 5.9$ and 9.4 , respectively) being rock-dominated, while Ganymede and Callisto (with $a/R_P = 15$ and 26.3) are a mixture of roughly half rock, half ice. We have proposed⁹ that these four satellites formed during the waning stages of inflow to Jupiter (with $\tau_G > 5 \times 10^6$ yr), and identified disk conditions consistent with the satellites' compositions, their survival against orbital decay, and a prolonged formation time required for Callisto's apparent state of incomplete differentiation^{2,42}. Saturn's system is dominated by Titan at (a/R_P) = 20.3, which contains ~96% of the total saturnian satellite system mass and is composed of about half rock, half ice. Between $3 < a/R_P < 9$ at Saturn are much smaller satellites containing a few per cent of Titan's mass, and having a variety of compositions, ranging from mixtures of rock and ice to nearly all ice⁴³. The uranian system is similar to that of Jupiter's in structure, but all four of its large satellites have densities implying roughly half-rock, half-ice compositions.

orbit. The torque is proportional to $(m_S/M_P)^2$, where m_S is the satellite's mass, while the satellite's orbital angular momentum is proportional to (m_S/M_P) . Thus density wave torques become more important as satellites grow.

Interactions with density waves circularize satellite orbits^{10–12}, and on a longer timescale, cause their orbital radii to decay through so-called type I migration^{13,14}. The associated eccentricity (e) and semi-major axis (a) decay timescales, $\tau_e = e/|\dot{e}|$ and $\tau_1 = a/|\dot{a}|$, are given by:

$$\tau_1 = \frac{1}{C_a \Omega} \left(\frac{M_P}{m_S} \right) \left(\frac{M_P}{r^2 \sigma_G} \right) \left(\frac{H}{r} \right)^2 = \frac{C_e}{C_a} \frac{\tau_e}{(H/r)^2} \quad (1)$$

where C_a and C_e are constants of order unity^{15,16}, σ_G is the disk gas surface density, $\Omega = (GM_P/r^3)^{1/2}$ is the keplerian angular velocity at radius r , and H is the vertical thickness of the gas with sound speed c , with $(H/r) \approx (c/r\Omega) \approx 0.1$.

In a disk supplied by an ongoing inflow, a satellite grows until it reaches a critical mass for which the characteristic time for its further growth is comparable to its type I orbital decay timescale. Satellites

cannot grow substantially larger than this critical mass before they are lost to collision with the planet. The accretion timescale for a mass m_S satellite is $\tau_{\text{acc}} \approx fm_S/(2\pi r \Delta r F_{\text{in}})$, where F_{in} is the inflow flux per area, f is the gas-to-solids mass ratio in the inflow, and $2\pi r \Delta r$ is the annular disk area over which the satellite accumulates material. We find (see Supplementary Notes) that the annulus width is a function of a satellite's characteristic maximum eccentricity, e , with $\Delta r/r \approx 2e$ and $e \approx (H/r)(m_S/4\pi r H \sigma_G)^{1/5}$, owing to a balance between eccentricity damping by density waves and excitation via gravitational scatterings with similarly sized objects¹⁷. Using such an estimate for Δr gives $\tau_{\text{acc}} \approx (f \sigma_G / F_{\text{in}})(m_S/4\pi r H \sigma_G)^{4/5}$.

The critical maximum satellite mass, m_{crit} , is then found by setting $\tau_{\text{acc}} = \tau_1$ from equation (1) and solving for $m_S = m_{\text{crit}}$. In quasi-steady state, the gas surface density σ_G is proportional to the inflow rate and inversely proportional to the rate at which the gas disk spreads, with the latter parameterized by a dimensionless constant¹⁸ α (so that $\sigma_G \propto (F_{\text{in}}/\alpha)$; see also Fig. 1 legend). We simplify by considering a uniform inflow flux per area across the disk interior to a radius r_C , so that $F_{\text{in}} \approx M_P/(\pi r_C^2 \tau_G)$, where $\tau_G \equiv M_P(dM/dt)^{-1}$ is

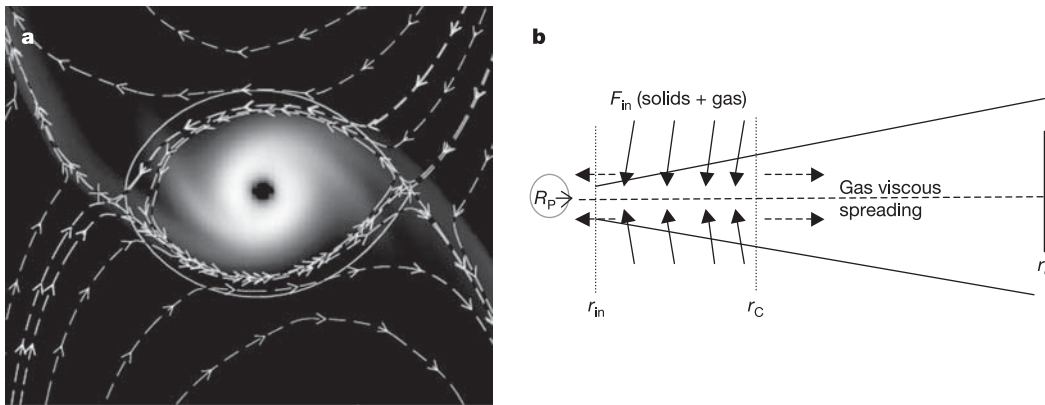


Figure 1 | An inflow-supplied circumplanetary disk. **a**, Hydrodynamical model⁶ of gas inflow into the Hill sphere (solid line) of a jovian-mass planet, viewed from above. Brightness scales with gas density; flow streamlines are dashed. **b**, The disk model considered here⁹, viewed edge-on. Gas and small solids from heliocentric orbits flow into the disk (indicated by solid arrows) across a region extending from an inner radius r_{in} to an outer radius, r_C , with an inflow flux per area $F_{\text{in}} \propto (1/r)^{\gamma_{\text{in}}}$ (where r is orbital radius and γ_{in} is an input parameter), with F_{in} containing a mass ratio f of gas-to-solids. Once in orbit, the gas viscously spreads (indicated by dashed arrows) with timescale $\tau_v \approx r^2/\nu$ and a viscosity¹⁸ $\nu = \alpha c H$, where $\alpha < 1$ is a parameter characterizing the strength of viscous turbulence in a gas disk of sound speed c and vertical scale height H . When $\tau_v \ll F_{\text{in}}/F_{\text{in}} \equiv \tau_{\text{in}}$, the gas maintains a quasi-steady-state surface density, σ_G (ref. 9). For $\gamma_{\text{in}} = 0$ and $r \leq r_C$, $\sigma_G \approx (0.3 F_{\text{in}} r_C^2 / \nu) [1.2 - \sqrt{r_C/r_d} - (r/r_C)^2/4] \approx 0.2 (1/\alpha) \times (F_{\text{in}} \Omega^{-1}) (r/H)^2 (r_C/r)^2$, with the bracketed quantity estimated for

$(r/r_C) = 0.5$ and $(r_C/r_d) = 0.2$ (ref. 9), where r_d is the disk outer edge and Ω is orbital frequency at radius r . Analytical estimates (see text) suggest that r_C is on the order of a few times $10R_P$, although because the planet's gravity alters the gas density and flow in its vicinity, hydrodynamical simulations are needed for improved estimates. Protostellar disk models⁵ consider $10^{-4} < \alpha < 0.1$, and $\tau_v \approx (1/\alpha)$ yr for protosatellite disks with $r \approx r_C \approx 30R_P$. While the source of viscosity in un-ionized disks is debated, several processes could contribute to protosatellite disks, including turbulence due to the velocity differential between inflowing and orbiting gas⁴⁴, baryclonic instability⁴⁵, and density wave torques from the satellites^{9,46}. Solar composition material has $f \approx 10^2$, while the bulk compositions of Jupiter and Saturn are in the range $f \approx 3$ to 30 (Table 1). During the satellite formation era, f could have been larger than solar if most solids were contained in large objects, or smaller if some of the gas nebula had already dissipated and fragmentation had maintained a supply of small material.

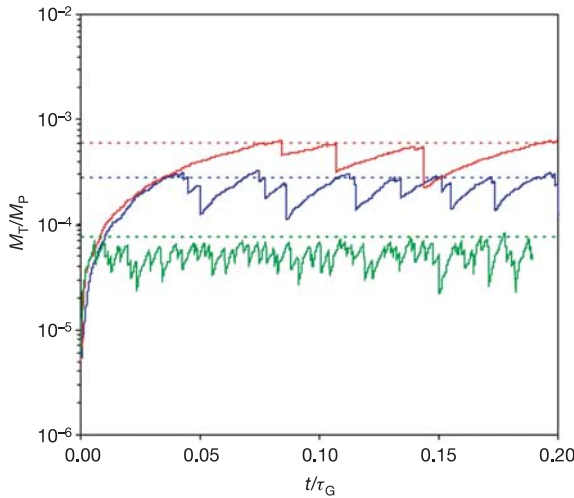


Figure 2 | Results of satellite accretion simulations with time-constant inflows. The total mass in satellites, M_T , scaled to the planet's mass, M_P , is shown versus time scaled to $\tau_G \equiv M_P / (dM/dt)^{-1}$, where dM/dt is the inflow rate. All three cases consider inflows having $\tau_G = 5 \times 10^6$ yr, $r_C = 30R_P$, and $\gamma_{in} = 0$, with the green, blue and red lines corresponding respectively to simulations with $(\alpha/f) = 10^{-6}$, 5×10^{-5} and 5×10^{-4} . The inflow of solids causes M_T to increase with time until objects of mass $\sim m_{crit}$ form (equation (2)). The orbits of the largest satellites then decay inward, and M_T decreases in discrete steps as satellites are lost to collision with the planet. Solid inflow to the disk continues, leading to the growth of another generation of mass $\sim m_{crit}$ objects, and the cycle repeats. As (M_T/M_P) depends on $(\alpha/f)^{1/3}$, the factor of 500 variation in (α/f) across these simulations yields about a factor of 10 spread in the characteristic system mass fractions. The long period oscillations in (M_T/M_P) reflect the time needed to deliver mass sufficient to form mass m_{crit} objects; this period shortens as (α/f) (and therefore (m_{crit}/M_P)) is decreased for a fixed τ_G . Shorter period variations result from the loss of individual objects. Dashed lines are predicted (M_T/M_P) values (equation (3)). Equations (2) and (3) treat disk annuli independently but in actuality, as satellites formed in the outer disk migrate inwards they pass through interior zones and cannibalize material along the way. Migration-driven growth hastens their orbital decay, so that they are lost somewhat more quickly than the time needed to replenish their mass in their original radial zone. This causes both a spread in the maximum satellite mass at a given time compared to equation (2), and the (M_T/M_P) value from equation (3) to be an approximate upper limit, as seen here.

the timescale for delivery of mass M_P . The critical satellite mass in planet masses is then (see also Supplementary Notes):

$$\left(\frac{m_{crit}}{M_P}\right) \approx 5.4 \left(\frac{\pi}{C_a}\right)^{5/9} \left(\frac{H}{r}\right)^{26/9} \left(\frac{r}{r_C}\right)^{10/9} \left(\frac{\alpha}{f}\right)^{2/3} (\Omega \tau_G f)^{1/9} \quad (2)$$

$$\approx 5.6 \times 10^{-5} \chi \left(\frac{3.5}{C_a}\right)^{5/9} \left(\frac{H/r}{0.1}\right)^{26/9} \left(\frac{r/r_C}{0.5}\right)^{10/9} \left(\frac{\alpha/f}{3 \times 10^{-5}}\right)^{2/3}$$

where $\chi \equiv [(1 \text{ week}/\{2\pi/\Omega\})(f/10^2)(\tau_G/10^7 \text{ yr})]^{1/9}$ is of order unity. The ratio (m_{crit}/M_P) depends extremely weakly on the inflow rate through the $(\tau_G)^{1/9}$ term in χ . Thus a similar maximum satellite mass would result for a wide range of inflow rates. The disk aspect ratio, (H/r) , is a slowly varying quantity with r for most disks, with $(H/r) \approx 0.1$ (ref. 9). Although r_C could potentially vary substantially between planets, so long as satellites form throughout the inflow region, the ratio (r/r_C) will be similar and of order unity for the largest satellites. The last term in equation (2) contains the ratio of two key parameters: the viscosity α parameter, and the gas-to-solids ratio in the inflow, f . For a given inflow flux, a higher viscosity yields lower disk gas surface densities, and thus allows larger mass satellites to survive against type I decay, while a lower f implies a more solid-rich inflow, which hastens the rate of satellite growth so that objects grow larger before they are lost.

We now consider the implications of this limiting mass for the total mass of the resulting satellite system. Consider an inflow that persists for a time exceeding that needed for a satellite of mass m_{crit} to form. Within a given annulus in the disk, a satellite grows to a mass $\sim m_{crit}$ before being lost to type I decay, but in a comparable timescale to its loss another similarly massive satellite grows in its place (because $\tau_1 \approx \tau_{acc}$). In this way, the disk is regulated to contain a total mass in satellites, M_T , comparable to a distribution of mass m_{crit} objects across the inflow region. For (H/r) and f that are approximately constant across the disk, the predicted satellite system mass fraction is:

$$\left(\frac{M_T}{M_P}\right) = \int_{R_P}^{r_C} \frac{(m_{crit}/M_P) dr}{\Delta r}$$

$$\approx 3.5 \left(\frac{1}{C_a}\right)^{4/9} \left(\frac{H}{r}\right)^{10/9} \left(\frac{\alpha}{f}\right)^{1/3} \frac{1}{(\Omega \tau_G f)^{1/9}} \quad (3)$$

$$\sim 2.5 \times 10^{-4} \frac{1}{\chi} \left(\frac{3.5}{C_a}\right)^{4/9} \left(\frac{H/r}{0.1}\right)^{10/9} \left(\frac{\alpha/f}{3 \times 10^{-5}}\right)^{1/3}$$

similar to the observed satellite systems. Here we assume $r_C \gg R_P$, where R_P is the planet's radius. Note that (M_T/M_P) is insensitive to inflow rate through χ , lacks a dependence on r_C , and depends quite weakly on (α/f) .

Simulation results

We model satellite growth and loss using a direct N -body accretion simulation¹⁹, modified to include interactions with a gas disk and ongoing mass inflow (Supplementary Methods). The solid inflow is mimicked by the addition of orbiting objects with random positions within the inflow region at a rate proportional to (F_{in}/f) . Collisions are treated as inelastic mergers.

Figure 2 shows results of three simulations involving a time-constant gas inflow rate but varied values for (α/f) . Type I orbital decay acts as a negative feedback on the total mass contained in the satellite system, causing (M_T/M_P) to oscillate about a value comparable to the equation (3) estimate. Figures 3 and 4 show results of exponentially decaying, time-dependent inflows. If the total mass in solids delivered to the disk is comparable or greater than M_T , one or more satellite systems described by equations (2) and (3) result. For example, a solar composition ($f \approx 10^2$) inflow that provided the last 10% of a planet's mass could yield approximately five satellite systems with $M_T/M_P = 2 \times 10^{-4}$.

Comparison with observed satellite systems

Our findings reveal a range of systems, within which the most basic properties of the jovian, saturnian and uranian satellite distributions are found (Fig. 3; Supplementary Figures). These three planets' varied masses and compositions imply that they would have probably processed very different amounts of material through their proto-satellite disks, and yet their satellite system mass ratios are nearly identical. The results here predict this commonality, as the balance between inflow-regulated satellite growth and gas-driven satellite loss causes the total satellite mass fraction to maintain a roughly constant value, which is nearly independent of both the inflow rate and its characteristic specific angular momentum (which, as shown below, determines r_C). Thus, for example, a Jupiter-like gas giant which acquired most of its mass through gas inflow from a strongly perturbed disk (Fig. 1a) is predicted to share a common satellite mass fraction with that of a much smaller uranian-sized planet, whose gas inflow history and gravitational effects on the local nebula would have been substantially different. This common fractional value is determined primarily by the ratio of two parameters describing the disk's viscosity (α) and the inflow's gas-to-solid composition (f). A weak dependence on this ratio, with $(M_T/M_P) \propto (\alpha/f)^{1/3}$, implies that satellite system masses similar to those of

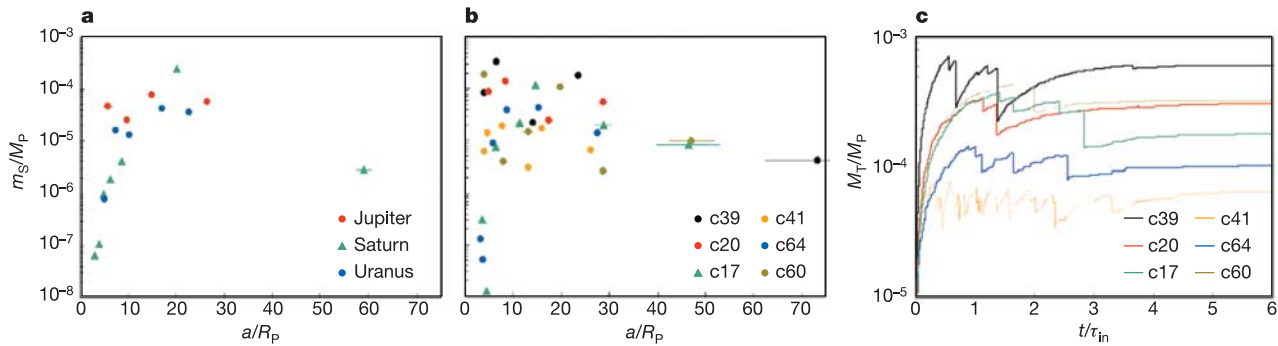


Figure 3 | Properties of observed satellites compared to simulations. Panel **a** shows observed satellites, while panels **b** and **c** contain final simulated systems produced by time-dependent inflows (with m_s , satellite mass; M_p planetary mass). Panels **b** and **c** show results of inflows with $\gamma_{in} = 0$, $r_C = 30R_p$, $1.7 < M_{in}/M_T < 10$ (where M_{in} is the total mass in solids delivered to the disk), and inflow exponential decay times, τ_{in} , between 2×10^5 yr and 1.5×10^6 yr. Systems were simulated for between 3×10^6 yr and 10^7 yr. Similar results are expected for higher M_{in}/M_T values (corresponding to longer decay times and/or higher initial inflow rates), provided that the corresponding inflow rate is consistent with a contracted planet and a circumplanetary disk. Horizontal lines connect periapse to apoapse for each satellite. **b**, A high, $(\alpha/f) = 5 \times 10^{-4}$ case, with $\alpha = 0.05$, produces five satellites with $(M_T/M_p) = 6.1 \times 10^{-4}$ (c39, black), while a low,

$(\alpha/f) = 10^{-6}$ case, with $\alpha = 10^{-4}$, produces six satellites containing $(M_T/M_p) = 6.6 \times 10^{-5}$ (c41, yellow). An $(\alpha/f) = 6.5 \times 10^{-5}$ case with $\alpha = 0.0065$ produces a galilean-like system with $(M_T/M_p) = 3.0 \times 10^{-4}$ (c20, red), while an $(\alpha/f) = 1.3 \times 10^{-5}$ case with $\alpha = 0.0065$ produces a uranian-like system with $(M_T/M_p) = 10^{-4}$ (c64, blue). A saturnian-like system with $(M_T/M_p) = 1.8 \times 10^{-4}$ (c17, green) results for $(\alpha/f) = 6 \times 10^{-5}$ and $\alpha = 0.006$. The most massive satellite at $14.6R_p$ contains $1.2 \times 10^{-4} M_p$ and 70% of the total satellite system mass, and its close orbital spacing to the satellite at $11.3R_p$ (with $C \approx 7$ in mutual Hill radii) suggests a future collision may occur²⁰ (Supplementary Notes). An $(\alpha/f) = 1.2 \times 10^{-4}$ case (c60, olive) yields $(M_T/M_p) = 3.3 \times 10^{-4}$, and two large satellites containing $0.9M_T$, with the innermost having nearly been lost to inward decay. **c**, (M_T/M_p) versus scaled time for the simulations shown in **b**.

Jupiter, Saturn and Uranus result for $10^{-6} < (\alpha/f) < 5 \times 10^{-4}$, that is, a span of nearly three orders of magnitude. Although reliable predictions for α and f are lacking, commonly used values (Fig. 1 legend) fall throughout this range.

The jovian and uranian systems each have four similarly sized large satellites, while Saturn has a single large satellite and numerous smaller satellites. Both morphologies can result from cycles of satellite formation and loss, depending on the timing of inflow cessation relative to the mass fraction oscillations seen in Fig. 2. As a system containing multiple, similarly sized satellites (for example, run c20 in Fig. 3b) undergoes orbital decay, it evolves through

periods in which only one or two of the largest satellites remain (for example, runs c17 and c60 in Fig. 3b), accompanied by much smaller moons that have accumulated in the regions vacated by the lost satellites, with this cycle repeating until the inflow ceases.

The median number of final satellites in our 75 simulated systems was $N = 7$. Future collisions on timescales longer than those simulated here are possible in many cases²⁰ (Supplementary Notes), which would generally reduce N . As our added objects are larger than those physically expected due to the inflow, the smallest final satellites resulting from only a few mergers were not well resolved. The median number of large satellites with $(m_s/M_p) \geq 10^{-5}$ was $N_{lg} = 4$. These objects had average orbital separations $\langle C_{lg} \rangle \approx 17$ comparable to both those of the outer planets (see Table 1) and our analytic estimates (Supplementary Notes).

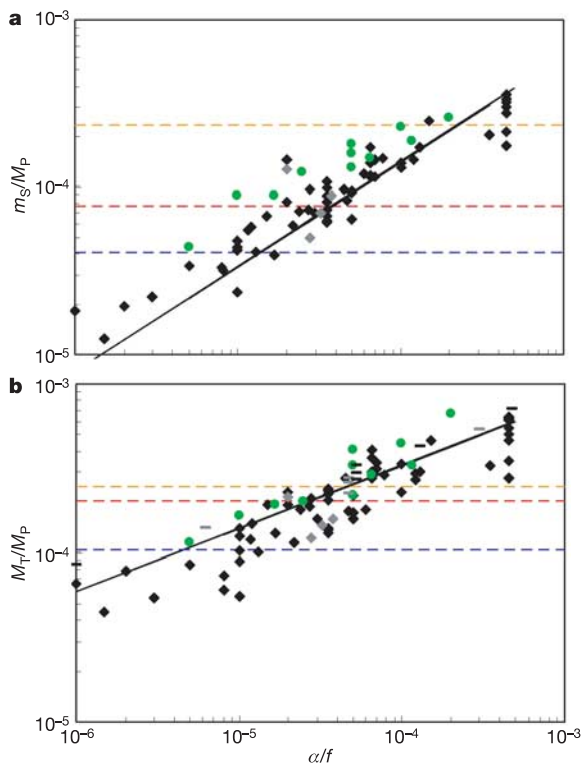


Figure 4 | Results of accretion simulations with time-dependent inflows. We consider $F_{in}(t) = F_{in}(0) \exp(-t/\tau_{in})$ and $\sigma_G(t) = \sigma_G(0) \exp(-t/\tau_{in})$, where τ_{in} is the inflow decay time (with $10^5 \leq \tau_{in}$ (years) $\leq 2 \times 10^6$) and $t = 0$ is when we begin to follow the system's evolution. Grey, black and green symbols correspond to cases with $(r_C/R_p) = 25, 30$ and 44 , respectively, while the red, orange and blue dashed lines show observed values for Jupiter, Saturn and Uranus. **a**, Maximum final satellite mass scaled to M_p versus (α/f) . (α , viscosity parameter; f , gas-to-solids ratio in the inflow.) **b**, Final satellite system total mass scaled to M_p versus (α/f) . Individual dashes are peak values from simulations with time-constant inflows (for example, Fig. 2). In **a** and **b**, black lines are equations (2) and (3) with $(r/r_C) = 0.5$, $(c/r\Omega) = 0.1$, $\tau_{in} = 10^6$ yr, and $\tau_G = \tau_{G,last} \approx \tau_{in}(M_p/M_T)/f$, corresponding to the last generation of satellites. The viscosity α values were estimated⁹ using an effective planet temperature $T_p = 500$ K and disk opacity $K = 0.1$; smaller (larger) values would result if T_p and/or K were higher (lower). By considering type I migration, we assume satellites do not grow massive enough to open annular gaps in the gas, at which point they transition to type II behaviour and are typically locked to the disk's viscous evolution^{14,47}. An estimate of the gap opening mass is⁴⁸ $m_{Gap}/M_p \approx C_v \sqrt{\alpha} (H/r)^{5/2}$, where C_v is a constant $\sim 1-10$. Our largest final satellites are all less massive than m_{Gap} for $C_v = 3$, with an average $\langle m_{lgst}/m_{Gap} \rangle = 0.2 \pm 0.1$. In an inflow-supplied disk, satellites are generally lost to type I decay before they grow to m_{Gap} . Type II behaviour would tend to accelerate orbital decay relative to type I (because $\tau_v \ll \tau_I$ for disks here), accentuating the limiting effects on M_T and m_s that we emphasize.

We find systems with maximum satellite orbital radii similar to those observed ($20 < a_{\max}/R_p < 60$) for an inflow outer radius (r_c) between 25 and 44 planetary radii (R_p), with up to a factor of two variation in $a_{\max}$ for fixed r_c owing to scattering and migration (Supplementary Fig. 2). This range for r_c is comparable to that implied by a three-dimensional estimate^{21,22} of the net specific angular momentum, j , of uniform density material entering a sphere of radius R_H (see below), neglecting the planet's gravity. This gives $j = \Omega_p R_H^2/5$ (where Ω_p is the planet's heliocentric orbital frequency, $R_H = a_p(M_p/3M_\star)^{1/3}$ is its Hill radius, a_p its semi-major axis, and $M_\star$ is the Sun's mass), implying a characteristic circumplanetary orbital radius $r_\star = j^2/GM_p$, with $r_\star/R_p \approx 10$ to 35 for Jupiter, Saturn and Uranus. A uniform inflow flux per area within r_c has $\langle j \rangle = \sqrt{GM_p r_\star}$ when $r_c \approx 1.6r_\star$.

Satellite compositions constrain the inflow properties during their growth, as the disk's effective temperature is a function of inflow rate (with $T_{\text{eff}} \propto F_{\text{in}}^{1/4}$, ref. 9). Thus early satellites forming during faster inflows (with smaller τ_c) would have masses set by equations (2) and (3) but rock-rich compositions (and a somewhat larger f)^{9,23}. As inflow slows, ice is increasingly incorporated. As an example, a final generation of jovian satellites produced by a solar composition inflow with a $\tau_{\text{in}} = 10^6$ -yr decay timescale (comparable to solar nebular lifetimes⁵) would form within a disk having temperatures below 200 K exterior to about $15R_p$ (with α approximately equal to a few $\times 10^{-3}$, a disk opacity²⁴ $K = O(10^{-1}) \text{ cm}^2 \text{ g}^{-1}$, and a planet temperature $T_p \approx 500$ K, ref. 9), consistent with Jupiter's outer icy moons. For a planet with a much lower mass and temperature³, a mixture of ice and rock consistent with the uranian satellite compositions is expected throughout most of the disk.

In our most Saturn-like systems (for example, c17 in Fig. 3b and Supplementary Fig. 1), one or more inner companions to the final 'Titan' collided with the planet. When an inner large satellite is lost as the inflow wanes, satellites containing a much smaller total mass that form in its place derive their material from cooler disk conditions than those which predominated during earlier growth. For a solar composition inflow with $\tau_{\text{in}} = 10^6$ yr, the last 10% of a satellite system with $(M_T/M_p) \approx O(10^{-4})$ is delivered at rates slow enough that the ice stability line moves inward to a few planetary radii for a Saturn-like planet. This would allow for ice incorporation in Saturn's small inner satellites.

Whereas our results offer strong support for a common mode of origin for the satellites of gas giants (like Jupiter and Saturn), the uranian system remains more uncertain. The uranian satellites' properties are consistent with those produced here, but the 98° tilt of Uranus' rotational axis requires additional explanation. A giant impact could have been responsible²⁵, with the current satellites forming subsequently. It is not obvious that post-impact gas inflow could yield retrograde satellites (that is, regular satellites of a retrograde spinning Uranus), although retrograde circumplanetary disks have resulted in some simulations of gas inflow to approximately Uranus-sized planets²⁶. The substantial $\sim 27^\circ$ obliquity of Saturn probably resulted from spin-orbit resonant interactions^{27,28}, which caused the planet's obliquity to change slowly enough that pre-existing regular satellites would have tracked its shift and remained aligned with the planet's equatorial plane²⁹. Recent work argues for a similarly gradual origin of Uranus' obliquity³⁰. Our results reveal an advantage in linking the uranian satellites to formation from an inflow-supplied rather than an impact-produced disk^{3,31,32}, in that the similarity in the system's mass fraction to those of Jupiter and Saturn would then be causal rather than coincidental. Further work incorporating the origin of the uranian obliquity is needed to ultimately distinguish between these alternatives.

Although the model presented here applies to regular satellite formation, it is remarkable that Neptune's single large, irregular satellite, Triton, also contains a similar mass fraction, with $M_{\text{Triton}}/M_p \approx 2.1 \times 10^{-4}$. Triton's orbit is retrograde and inclined, and it is believed to have been captured intact from heliocentric

orbit^{33–35}. Although we lack direct evidence of a putative^{33,35,36} original prograde neptunian satellite system, it seems clear that it cannot have contained much greater total mass than Triton itself. Otherwise, a retrograde and initially eccentric Triton would have been destroyed while traversing the regular satellite region (either as it was accreted or collisionally disrupted, or by the decay of its orbital angular momentum³⁶ owing to interactions with a much greater mass in prograde material). The survival of a captured retrograde satellite requires it to have comparable or greater mass than any prograde system with which it actively interacts. As larger interlopers would have been less numerous (and therefore captured less frequently), the most probable surviving Triton-like object would be one having the smallest mass affording its survival, which would suggest that M_{Triton} is similar to the total mass of Neptune's original regular satellites, or $M_T/M_{\text{Neptune}} \approx O(10^{-4})$ as well.

General implications

When more than a few per cent of a gas planet's mass in solar composition material is processed through a circumplanetary disk, one or more generations of inflow-produced satellite systems are likely^{9,23}, with earlier satellites doomed to collision with the planet. Today's observed satellites are then the last generations that formed as inflow to the planets waned, implying that they formed very slowly in low-pressure, 'gas-starved'³⁷ disks. For extrasolar giant planets orbiting within their stellar habitable zone, the prospect has been raised of the existence of habitable environments on hypothesized Earth-sized satellites^{38,39}. Provided that such planets accumulated gas while contracted to scales consistent with circumplanetary disks (which appears likely for nebular removal timescales $\geq 10^6$ years^{8,40}; Supplementary Notes), the findings here imply that their largest surviving satellites would contain on the order of $10^{-4}M_p$, so that a Jupiter-mass exoplanet would host only Moon-to-Mars sized satellites.

Received 13 March; accepted 5 May 2006.

- Jewitt, D. & Sheppard, S. Irregular satellites in the context of planet formation. *Space Sci. Rev.* **116**, 441–455 (2005).
- Stevenson, D. J., Harris, A. W. & Lunine, J. I. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) 39–88 (University of Arizona Press, Tucson, 1986).
- Pollack, J. B., Lunine, J. I. & Tittlemore, W. C. in *Uranus* (eds Bergstrahl, J. T., Miner, E. D. & Matthews, M. S.) 469–512 (University of Arizona Press, Tucson, 1991).
- Lunine, J. I., Coradini, A., Gautier, D., Owen, T. C. & Wuchterl, G. in *Jupiter: The Planet, Satellites and Magnetosphere* (eds Bagenal, F., Dowling, T. E. & McKinnon, W. B.) 19–34 (Cambridge University Press, Cambridge, UK, 2004).
- Calvet, N., Hartmann, L. & Strom, S. E. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 377–399 (University of Arizona Press, Tucson, 2000).
- Lubow, S. H., Seibert, M. & Artymowicz, P. Disk accretion onto high-mass planets. *Astrophys. J.* **526**, 1001–1012 (1999).
- Bate, M. R., Lubow, S. H., Ogilvie, G. I. & Miller, K. A. Three-dimensional calculations of high- and low-mass planets embedded in protoplanetary discs. *Mon. Not. R. Astron. Soc.* **341**, 213–229 (2003).
- Papaloizou, J. C. B. & Nelson, R. P. Models of accreting gas giant protoplanets in protostellar disks. *Astron. Astrophys.* **433**, 247–265 (2005).
- Canup, R. M. & Ward, W. R. Formation of the Galilean satellites: Conditions of accretion. *Astron. J.* **124**, 3404–3423 (2002).
- Ward, W. R. On disk-planet interactions and orbital eccentricities. *Icarus* **73**, 330–348 (1998).
- Artymowicz, P. Disk-satellite interaction via density waves and the eccentricity evolution of bodies embedded in disks. *Astrophys. J.* **419**, 166–180 (1993).
- Papaloizou, J. C. B. & Larwood, J. D. On the orbital evolution and growth of protoplanets embedded in a gaseous disc. *Mon. Not. R. Astron. Soc.* **315**, 823–833 (2000).
- Ward, W. R. Density waves in the solar nebula—Differential Lindblad torque. *Icarus* **67**, 164–180 (1986).
- Ward, W. R. Protoplanet migration by nebula tides. *Icarus* **126**, 261–281 (1997).
- Tanaka, H. & Ward, W. R. Three-dimensional interaction between a planet and an isothermal gaseous disk. II. Eccentricity waves and bending waves. *Astrophys. J.* **602**, 388–395 (2004).
- Tanaka, H., Takeuchi, T. & Ward, W. R. Three-dimensional interaction between a planet and an isothermal gaseous disk. I. Corotation and Lindblad torques and planet migration. *Astrophys. J.* **565**, 1257–1274 (2002).

17. Ward, W. R. Density waves in the solar nebula—Planetesimal velocities. *Icarus* **106**, 274–287 (1993).
18. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24**, 337–355 (1973).
19. Duncan, M. J., Levison, H. F. & Lee, M. H. A multiple timestep symplectic algorithm for integrating close encounters. *Astron. J.* **116**, 2067–2077 (1998).
20. Chambers, J. E., Wetherill, G. W. & Boss, A. P. The stability of multi-planet systems. *Icarus* **119**, 261–268 (1996).
21. Lissauer, J. J. Urey Prize Lecture: On the diversity of plausible planetary systems. *Icarus* **114**, 217–236 (1995).
22. Mosqueira, I. & Estrada, P. R. Formation of regular satellites of giant planets in an extended gaseous nebula I: Subnebula model and accretion of satellites. *Icarus* **163**, 198–231 (2003).
23. Alibert, Y., Mousis, O. & Benz, W. Modeling the Jovian subnebula. I. Thermodynamic conditions and migration of proto-satellites. *Astron. Astrophys.* **439**, 1205–1213 (2005).
24. Hubickyj, O., Bodenheimer, P. & Lissauer, J. J. Accretion of the gaseous envelope of Jupiter around a 5 to 10 Earth-mass core. *Icarus* **179**, 415–431 (2005).
25. Slattery, W. L., Benz, W. & Cameron, A. G. W. Giant impacts on a primitive Uranus. *Icarus* **99**, 167–174 (1992).
26. D'Angelo, G., Henning, T. & Kley, W. Thermodynamics of circumstellar disks with high-mass planets. *Astrophys. J.* **599**, 548–576 (2003).
27. Ward, W. R. & Hamilton, D. P. Tilting Saturn. I. Analytic model. *Astron. J.* **128**, 2501–2519 (2004).
28. Hamilton, D. P. & Ward, W. R. Tilting Saturn II. Numerical model. *Astron. J.* **128**, 2510–2517 (2004).
29. Goldreich, P. Inclination of satellite orbits about an oblate precessing planet. *Astron. J.* **70**, 5–9 (1965).
30. Brunini, A. Origin of the obliquities of the giant planets in mutual interactions in the early Solar System. *Nature* **440**, 1163–1165 (2006).
31. Canup, R. M. & Ward, W. R. A possible impact origin of the Uranian satellite system. *Bull. Am. Astron. Soc.* **32**, 1105 (2000).
32. Ward, W. R. & Canup, R. M. Viscous evolution of an impact generated disk around Uranus. *Bull. Am. Astron. Soc.* **35**, 1046 (2003).
33. Goldreich, P., Murray, N., Longaretti, P. Y. & Banfield, D. Neptune's story. *Science* **245**, 500–504 (1989).
34. McKinnon, W. B., Lunine, J. I. & Banfield, D. in *Neptune and Triton* (ed. Cruikshank, D. P.) 807–877 (University of Arizona Press, Tucson, 1995).
35. Agnor, C. B. & Hamilton, D. P. Neptune's capture of its moon Triton in a binary-planet gravitational encounter. *Nature* **441**, 192–194 (2006).
36. Cuk, M. & Gladman, B. J. Constraints on the orbital evolution of Triton. *Astrophys. J.* **626**, L113–L116 (2005).
37. Stevenson, D. J. Jupiter and its moons. *Science* **294**, 71–72 (2001).
38. Williams, D. M. & Kasting, J. F. Habitable moons around extrasolar giant planets. *Nature* **385**, 234–236 (1997).
39. Barnes, J. W. & O'Brien, D. P. Stability of satellites around close-in extrasolar giant planets. *Astrophys. J.* **575**, 1087–1093 (2002).
40. Burrows, A., Hubbard, W. B. & Lunine, J. K. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* **73**, 719–765 (2001).
41. Saumon, D. & Guillot, T. Shock compression of deuterium and the interiors of Jupiter and Saturn. *Astrophys. J.* **609**, 1170–1180 (2004).
42. Anderson, J. D. *et al.* Shape, mean radius, gravity field, and interior structure of Callisto. *Icarus* **153**, 157–161 (2001).
43. Johnson, T. V. & Lunine, J. I. Saturn's moon Phoebe as a captured body from the outer Solar System. *Nature* **435**, 69–71 (2005).
44. Cassen, P. & Moosman, A. On the formation of protostellar disks. *Icarus* **48**, 353–376 (1981).
45. Klahr, H. H. & Bodenheimer, P. Turbulence in accretion disks: vorticity generation and angular momentum transport via the global baroclinic instability. *Astrophys. J.* **582**, 869–892 (2003).
46. Goodman, J. & Rafikov, R. R. Planetary torques as the viscosity of protoplanetary disks. *Astrophys. J.* **552**, 793–802 (2001).
47. Lin, D. N. C. & Papaloizou, J. C. B. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 749–835 (University of Arizona Press, Tucson, 1993).
48. Ward, W. R. & Hahn, J. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 1135–1155 (University of Arizona Press, Tucson, 2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by NASA's Planetary Geology and Geophysics and Outer Planets Research programmes. Computer resources and support were provided by SwRI. We thank L. Dones, R. Mhuran and S. Peale for comments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.M.C. (robin@boulder.swri.edu).

ARTICLES

Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise

John R. S. Newman^{1,2}, Sina Ghaemmamhami^{1,2,†}, Jan Ihmels^{1,2}, David K. Breslow^{1,2}, Matthew Noble¹, Joseph L. DeRisi^{1,3} & Jonathan S. Weissman^{1,2}

A major goal of biology is to provide a quantitative description of cellular behaviour. This task, however, has been hampered by the difficulty in measuring protein abundances and their variation. Here we present a strategy that pairs high-throughput flow cytometry and a library of GFP-tagged yeast strains to monitor rapidly and precisely protein levels at single-cell resolution. Bulk protein abundance measurements of >2,500 proteins in rich and minimal media provide a detailed view of the cellular response to these conditions, and capture many changes not observed by DNA microarray analyses. Our single-cell data argue that noise in protein expression is dominated by the stochastic production/destruction of messenger RNAs. Beyond this global trend, there are dramatic protein-specific differences in noise that are strongly correlated with a protein's mode of transcription and its function. For example, proteins that respond to environmental changes are noisy whereas those involved in protein synthesis are quiet. Thus, these studies reveal a remarkable structure to biological noise and suggest that protein noise levels have been selected to reflect the costs and potential benefits of this variation.

Proteins carry out the actions necessary for cellular function. Correspondingly, changes in their concentrations can have direct phenotypic consequences. Several approaches have been used to identify proteins expressed in a diverse range of organisms and organelles^{1–4}. In particular, studies on the budding yeast *Saccharomyces cerevisiae* have generated a nearly comprehensive list of proteins likely to be translated^{5–7} and their copy numbers at steady state in rich medium⁶. However, much less quantitative information is available on how proteomes are remodelled in response to different environmental cues. Additionally, inferences about changes in protein levels based on DNA microarray measurements are limited by an imperfect understanding of the relationship between mRNA and protein levels^{6,8–10}.

Notably, no widely used proteomics technique readily provides abundance measurements for single cells, although such information is critical for understanding fundamental biological questions. Indeed, it has long been appreciated that many cellular processes rely on small numbers of molecules and thus are subject to stochastic variation (noise)¹¹. Such noise contributes to phenotypic variability and can either attenuate or augment a cell's ability to respond to its environment^{12–14}. Even in the absence of noise, epigenetic imprints¹⁵ and a lack of synchronicity within a population can differentially impact transcription from genetically identical loci. Also, protein changes can occur in both graded and binary (switch-like) fashions in response to differing environmental conditions, but bulk measurements obscure such responses^{16–18}. Thus, identifying single-cell variation is essential for understanding how cells exist as autonomously functioning dynamic systems¹⁹.

Pioneering studies have highlighted the potential of flow cytometry to quantify intracellular protein concentrations in individual cells²⁰,

and to measure different samples rapidly²¹. More recently, several elegant studies using the green fluorescent protein (GFP) and its derivatives have identified many factors that can contribute to noise^{15,22–25}. However, the scope and nature of the central factors that give rise to and limit noise *in vivo* remain poorly understood (see ref. 26), in large part because existing studies have examined only a handful of genes.

Here we describe a novel, integrated strategy for large-scale, quantitative single-cell proteomics. Our approach uses high-throughput flow cytometry to make precise measurements on a collection of *S. cerevisiae* strains in which each protein is expressed as a carboxy-terminal GFP fusion from its endogenous promoter and natural chromosomal position³. With this approach, it is now possible to monitor the internal state of the cell with unprecedented resolution. This allows us to follow protein changes in response to environmental perturbations and to uncover both the global and protein-specific biological structure of noise in budding yeast.

Measuring protein abundance by flow cytometry

To quantify the yeast proteome, we automated sample growth and handling, and also wrote custom software (HTS-Pro; see Supplementary Notes 1 and 2) to control the delivery of GFP-tagged cells to a flow cytometer via an autosampling device. With this approach, approximately seven samples can be measured per minute, counting >50,000 cells per sample, with cross-contamination of <0.1%. Furthermore, we developed an integrated set of software tools that allows storage, manipulation and analysis of data (for example, to eliminate systematic sources of variation, and to extract information about cell size, shape and fluorescence; see Supplementary Notes 2).

We measured the fluorescence of 4,159 GFP-tagged yeast strains

¹Howard Hughes Medical Institute, ²Department of Cellular and Molecular Pharmacology, and ³Department of Biochemistry and Biophysics, University of California, San Francisco, and the California Institute for Quantitative Biomedical Research, 1700 4th Street, San Francisco, California 94107, USA. †Present address: Institute for Neurodegenerative Diseases, PO Box 0518, University of California, San Francisco, San Francisco, California 94143-0518, USA.

grown to mid-log phase in rich (YEPD) medium, starting from single colonies. These strains define the set that can be detected by fluorescence microscopy under normal laboratory growth conditions³. The flow cytometry measurements show remarkable precision and sensitivity. The fluorescence is highly reproducible (coefficient of correlation, $r^2 = 0.99$), even when strains are grown in medium prepared on different occasions and the measurements are separated by several weeks (Fig. 1a). The very low variability seen in full biological replicates probably results from instrument accuracy and from the minimal sample preparation required (that is, measurements are made on intact cells)²⁷. The sensitivity of the approach permits us to distinguish $>60\%$ ($>2,500$) of the GFP-tagged strains from autofluorescence with $>95\%$ confidence (Supplementary Tables S1 and S3). After correcting for autofluorescence, the standard errors for the lowest GFP fluorescence intensities are still less than $\pm 45\%$ of the median values. For larger fluorescence intensities, the standard errors fall to about $\pm 2.5\%$ of the median values (see Supplementary Notes 2).

Comparison of the proteins we detect to quantitative measurements obtained under similar growth conditions using western blot analysis of a set of tandem affinity purification (TAP)-tagged strains⁶ reveals that cytometry detects nearly all proteins present at greater than $\sim 8,000$ tagged proteins per cell. This coverage drops to about 50% for proteins expressed at 2,000–4,000 copies per cell, and falls

rapidly thereafter (Fig. 1b). However, we estimate that moderate improvements in fluorescent proteins^{28,29}, coupled to lower autofluorescence and reduced instrument drift, would permit us to detect $\sim 85\%$ of the proteome (data not shown). Alternatively, microscopy can be used to quantify proteins with low abundances^{3,30} (Supplementary Fig. S6a; J.R.S.N. and J.S.W., unpublished observation).

Several lines of evidence argue that our measured fluorescence accurately reports on protein abundance. First, the fluorescence does not change substantially on the timescale required to harvest and measure cell fluorescence (Supplementary Fig. S5). Second, the fluorescence measurements are in excellent agreement ($r^2 = 0.997$) with quantitative fluorescence microscopy (Supplementary Fig. S6a). Third, with the exception of proteins localized to the vacuole, cellular fluorescence is almost exclusively due to intact fusion proteins (rather than cleaved GFP), as confirmed by western blotting of >150 strains (Supplementary Fig. S7c). Fourth, there are no global problems with the folding and oxidation of GFP, as the fluorescence measured by cytometry is linearly proportional to abundances determined by western blotting using TAP-tagged strains ($r^2 = 0.80$; Supplementary Fig. S7a). The degree of agreement between the two approaches is consistent with errors from duplicate western blotting measurements ($r^2 = 0.77$; Supplementary Fig. S7b; see also ref. 30). Finally, we found no evidence that GFP interferes with the recognition and destruction of proteins by the ubiquitin–proteasome system, which

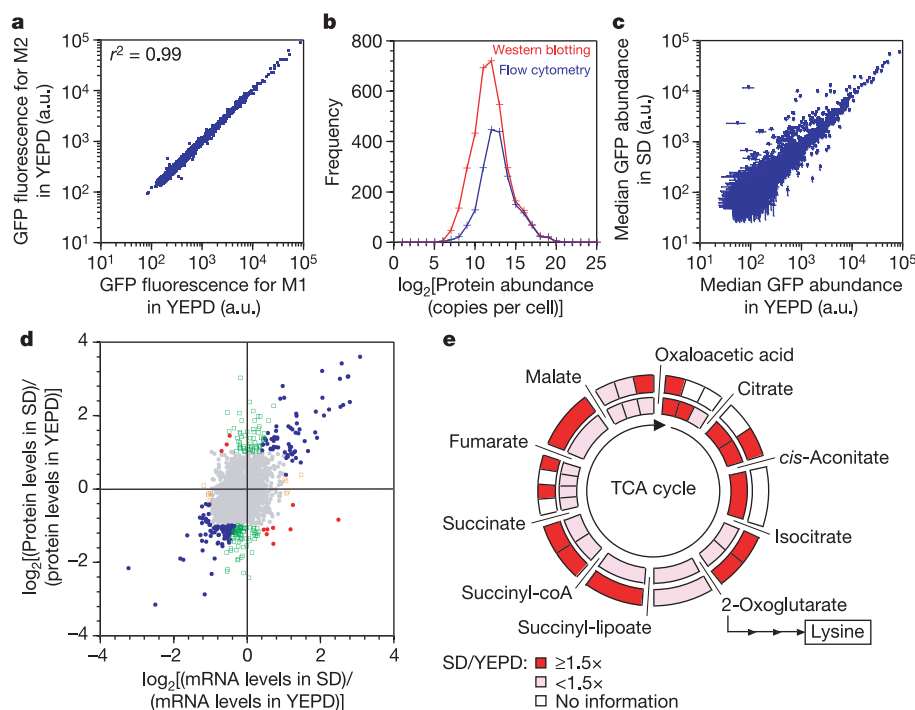


Figure 1 | Quantitative analyses of protein abundance using flow cytometry. **a**, Median fluorescence values for biological replicates of $\sim 4,159$ GFP-tagged strains grown in rich (YEPD) medium plotted against each other. M1, measurement 1; M2, measurement 2; a.u., arbitrary units. **b**, The frequency of detected, tagged strains is plotted as a function of the number of protein copies per cell ($\log_2$). TAP-tagged strains (red) were detected by western blotting⁶; GFP-tagged strains (blue) were detected by cytometry. The former approach detects essentially all tagged proteins present at more than 50 copies per cell, and thus provides a benchmark for evaluating the sensitivity of new approaches. **c**, Protein abundance measurements for 2,223 strains grown in rich (YEPD) and minimal (SD + Leu + Met + Ura) media. The standard errors of the measurements are also shown. Fluorescence from 313 strains can be quantified in YEPD but not SD, and 235 strains can be quantified in SD but not YEPD (data not shown). However, 2,763 strains (66% of the GFP library) can be quantified in at least one condition. **d**, The relationship between protein and mRNA

ratios larger than two (colours other than grey) for cells grown in SD and YEPD. Changes in mRNA levels are largely captured by changes in protein levels (blue). In 21 cases, mRNA levels change without a corresponding change in protein levels. For 10 of these cases, protein levels do not change (orange), whereas for the remaining 11 cases (red), the proteins change in a direction opposite to that observed for the mRNA (see the main text). In contrast, changes in protein levels are not always captured by changes in mRNA levels, and we observed 131 instances of this behaviour (green). mRNA ratios were measured using DNA microarrays. Ratios were grouped operationally by lines having slopes of $+3/3$ (blue), $-3/-3$ (red), $+3/-3$ (green) and $1/3-1/3$ (orange). **e**, Schematic showing that 11 out of 14 tricarboxylic acid (TCA) cycle proteins quantified by cytometry show greater than 1.5-fold induction in SD compared to YEPD (red segments, outer circle), but only 4 out of 19 mRNAs quantified by microarray analysis show induction at a similar level (red segments, inner circle). Note that Aco1 is used twice in the TCA cycle.

is a major mechanism for protein turnover³¹ (Supplementary Fig. S8; D. Toczyski, personal communication). Although, for a subset of proteins, the presence of a GFP tag will alter abundances, the large majority of fusion proteins (>80% based on analysis of essential

proteins³) remains functional. Notably, even when the GFP tag influences abundances, relative changes can still be determined accurately, provided that the tag perturbs protein levels by a constant amount across different conditions.

Protein and mRNA levels in rich and minimal media

Next, we explored the differences between steady-state protein levels for cells grown in rich (YEPD) and minimal (SD) media (Fig. 1c; Supplementary Table S1). Of the 2,223 proteins that can be detected in both conditions, 232 are expressed at higher levels in SD (>99% confidence), and 101 increase by at least twofold. Conversely, 684 proteins are expressed at higher levels in YEPD (>99% confidence), and 112 are induced by at least twofold. These changes provide a coherent description of how cells adapt to environments with abundant or modest levels of nutrients consistent with our understanding of this biological response^{10,32}. For example, many proteins induced in rich medium are involved in cell growth and division (such as ribosomal proteins and cell wall biosynthesis enzymes). Conversely, many proteins induced in minimal medium are involved in the production of small molecules that cannot be taken up from the surrounding environment, such as amino acids and nucleotides.

To understand how protein and mRNA changes are related, and to identify potential examples of post-transcriptional regulation (PTR; that is, regulation at the level of translation or protein degradation), we extended our studies by using DNA microarrays to quantify fluctuations in mRNA levels between cells grown in YEPD versus SD (Supplementary Table S2). Overall, we find that mRNA changes are largely captured by changes in protein abundances (Fig. 1d). For example, focusing on genes for which mRNA levels change at least twofold, we find 151 instances where both mRNA and protein level changes occur in the same direction, and only 21 instances where mRNA levels change and protein levels remain static or change in the opposite direction (see Fig. 1d legend). In fact, further studies using quantitative polymerase chain reaction (qPCR) revealed that several of these discrepancies were due to errors in the microarray measurements (data not shown). It is important to note that, as mRNA measurements are made using untagged strains, these results also argue that the GFP tag does not significantly impair our ability to detect physiological responses.

In contrast, there are a significant number of cases (135, using a twofold threshold) where protein changes are not mirrored by mRNA changes (Fig. 1d). Many of these protein changes yield insights into coordinated cellular responses that occur when cells are grown to steady state in YEPD or SD. One qualitative illustration is provided by the induction of tricarboxylic acid (TCA) cycle enzymes in SD, probably to generate key intermediates involved in amino acid biosynthesis (such as 2-oxoglutarate for lysine biosynthesis). Here, 11 out of 14 measured proteins are coordinately upregulated >1.5-fold (outer circle in Fig. 1e), whereas only 4 out of 19 measured genes are coordinately upregulated to the same extent (inner circle in Fig. 1e). Additional qualitative examples are provided by proteins and genes involved in 20S and 35S ribosomal RNA processing (19/24 proteins are induced in YEPD versus 4/24 mRNAs, and 13/24 proteins are induced in YEPD versus 3/24 mRNAs, respectively; Supplementary Table S2).

More quantitatively, we sought to independently verify examples of PTR. Because rare examples of PTR can be masked by measurement inaccuracies, particularly when the total number of measurements is large, the ability to identify infrequent cases of PTR provides a critical test of the overall accuracy of an approach. For our studies, we chose ten cases where additional microarray measurements performed in triplicate confirmed that mRNA levels change discordantly from protein levels when cells are shifted from SD to YEPD. Next, we used western blotting to confirm the protein measurements made by cytometry, and qPCR to confirm the mRNA measurements made by microarrays. In all cases examined, we were able to corroborate the discordance between protein and mRNA changes

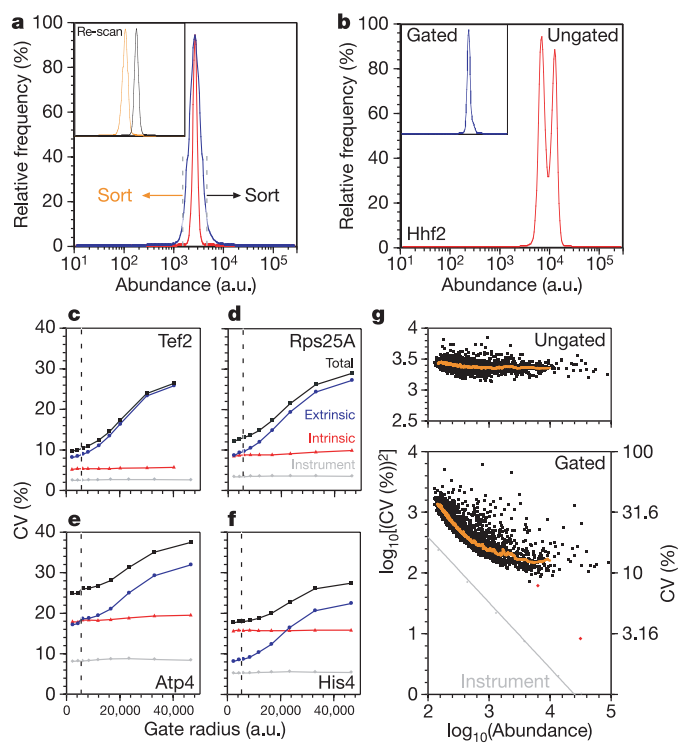


Figure 2 | Single-cell variation, gating and global trends in noise. **a**, Two GFP-tagged strains with the same mean intensities show different degrees of variation. Rpl35A–GFP has a CV of 11.8 (red), whereas Nop8–GFP has a CV of 38.6 (blue). Cells from the same population with low or high intensities retain their original fluorescence intensities when sorted and re-measured (inset), indicating that variation is a reproducible cellular property. The scales of the main and inset plots are identical (omitted for clarity; true also for **b**). **b**, Cells expressing the cell-cycle-regulated histone subunit Hhf2 fused to GFP normally show two peaks (red) in asynchronous cultures grown to mid-log phase in YEPD. Gated cells show a single peak, and a lower total variation (inset, blue). The gated peak can be exactly superimposed on the left peak in the ungated population (not shown for clarity). **c–f**, Decomposition of total noise (black) into intrinsic (red) and extrinsic (blue) contributions using two-colour strains and different FSC and SSC gate radii. Intrinsic noise due to the instrument for the voltages used for each experiment is also shown (grey). The gate radius used for CV calculations reported in the paper (4,096 a.u.) is indicated by a vertical dotted line in each graph. For gated cells, extrinsic noise dominates the total noise of the abundant protein Tef2 (**c**; abundance = 30,700). Intrinsic and extrinsic noise make equal contributions to the total noise of the ribosomal protein Rps25A (**d**; abundance = 6,132). Intrinsic and extrinsic noise make equal contributions to the total noise of Atp4, but intrinsic noise is consistently high (**e**; abundance = 1,276). For gated cells, the intrinsic noise of His4 dominates the total noise (**f**; abundance = 2,090). **g**, Global noise trends are revealed by gating. There is little dependence of $\log_{10}(\text{CV}^2)$ on $\log_{10}(\text{abundance})$ for strains gated using a large radius (40,960 a.u.; top panel). In contrast, there is a striking reduction in $\log_{10}(\text{CV}^2)$ with increasing abundance when a radius chosen to minimize extrinsic noise is used (4,096 a.u.; bottom panel). The r^2 for the correlation between CV^2 and $1/\text{abundance}$ is 0.64 (see Supplementary Notes 2). Uncorrelated stochastic processes are expected to show a slope of -1 . This relationship is observed for proteins expressed at low and moderate levels. For highly expressed proteins, extrinsic noise dominates the measured variation, but this component can be eliminated using a two-colour strategy²⁴ to reveal underlying noise consistent with stochastic processes (for example, for Rps25A–GFP and Tef2–GFP; red diamonds). A running median for each plot is also shown (orange). Instrument noise is shown in the bottom panel (grey line).

(see Supplementary Notes 2; Supplementary Fig. S9). Further validation of two of our observations (involving Fet3 and Ftr1) is provided by a recent report showing that iron uptake by these proteins is controlled post-transcriptionally by regulated proteolysis³³.

Defining global trends in cellular noise

A central difference between this study and previous large-scale analyses of protein or mRNA abundances is that cytometry functions with single-cell resolution and thus can report on cell-to-cell variation. Indeed, we see large differences in coefficient of variation (CV; (standard deviation/mean) $\times 100(\%)$) values for proteins expressed at similar levels (Fig. 2a). These differences are reproducible, as cells with low and high fluorescence intensities can be separated from a single population by sorting, and retain their original fluorescence levels when re-measured immediately (Fig. 2a, inset).

Extracting biological or functional meaning from single-cell data can be challenging. Specifically, there can be large global differences in protein levels between isogenic cells due to heterogeneity in cell size and cell cycle state that obscure protein-specific variation (data not shown and Fig. 2b, respectively)^{34,35}. In principle, cytometry is well suited to reducing the contributions of these factors to protein variation as it can provide an approximate measure of cell size and granularity (using the forward (FSC) and side (SSC) scatter parameters, respectively). Nevertheless, the effectiveness and importance of correcting for such effects is unclear^{34,35}. Therefore, we systematically explored the ability of gating to extract biological structure from global abundance heterogeneity by defining a circular gate centred about the FSC and SSC medians and varying its radius. Reducing the radius initially led to a steep decrease, and then a more moderate decline, in measured CV values (Fig. 2c–f; Supplementary Fig. S3a). For our studies, we used a radius that fell in the latter regime, chosen to achieve a substantial reduction in variation, while maintaining

sufficient numbers of cells (that is, ~ 500 , or 1% of all cells measured in a typical experiment) to reproducibly calculate CV values: 99% of the CV values from gated populations fall within $\pm 20\%$ of the average of repeated measurements, and this reproducibility is maintained with increasing CV magnitude (see Supplementary Notes 2; Supplementary Fig. S3d, e). Although, following gating, protein variation is typically well approximated by a normal distribution, a notable subset of tagged proteins shows distributions with significant tails, implying that our reported CV values sometimes underestimate the variation at the extremes of populations. Such deviations from normal distributions, which will be described in detail elsewhere, suggest that there may be multiple processes acting in concert to generate noise.

The critical role of gating in isolating cells with uniform properties is revealed by two observations. First, gating eliminates one of the two peaks normally observed for the GFP-tagged, cell-cycle-regulated, histone subunit Hhf2 (Fig. 2b, inset; see also Supplementary Notes 2). Second, gating leads to a dramatic reduction in noise due to global differences between cells, as revealed using a modification of the two-colour experiment described by Elowitz *et al.*²⁴ that permits differentiation between correlated (extrinsic) noise arising from variation between cells, and uncorrelated (intrinsic) noise that arises from stochasticity in gene expression/protein production (see Supplementary Notes 2). Using four different double-tagged diploid yeast strains (that is, with one chromosomal locus of an open reading frame tagged with GFP, and the second with a red fluorescent protein variant (tdTomato²⁹)), we found that whereas ungated populations are dominated by extrinsic noise, reduction of the FSC and SSC gate radii decreased the contribution of extrinsic noise either to levels comparable to intrinsic noise (Tef2, Atp4 and Rps25A), or, in one case (His4), to a level below that of intrinsic noise (Fig. 2c–f; see also Supplementary Notes 2). As expected, gating has a minimal effect on the levels of intrinsic noise.

These two-colour studies indicate that intrinsic noise makes a significant contribution to the total variation measured for gated strains. In turn, this predicts that our data should exhibit the characteristic inverse-proportional relationship between CV² and protein abundance (ref. 36) associated with fluctuations due to uncorrelated stochastic processes, such as the production and destruction of protein or mRNA molecules. Indeed, a strong inverse proportionality is observed for gated (but not ungated) measurements, particularly for proteins of low and medium abundances (Fig. 2g). Notably, this is not due to the response of the cytometer, which we measured directly (see Supplementary Notes 2). In contrast, for abundant proteins we find that concentration-independent noise due to cellular heterogeneity dominates. When this extrinsic noise is excluded for two highly expressed proteins, Tef2 and Rps25A (studied using the two-colour approach described above), the intrinsic noise is found to be close to that predicted by extrapolation from the linear dependence of noise seen for lower abundance proteins (Fig. 2g).

Biological structure in protein variation

Beyond these global relationships between noise and protein abundance, we observe substantial differences in variation on a protein-by-protein basis (for example, see Fig. 2g). To extract and interpret this information in a systematic manner, we calculated the distance of each CV to a running median of CV values (hereafter referred to as DM; Supplementary Table S1). DM values permit protein-specific noise levels to be compared independently of confounding influences due to protein abundance, noise from the instrument response, or intracellular differences in cells globally affecting protein production.

We looked for correlations between DM values and both known and potential factors that might influence noise, including those that act at multiple points during protein production, organization and destruction (Figs 3 and 4; see also Discussion and Supplementary Notes 2). We find that a number of factors shown or postulated to

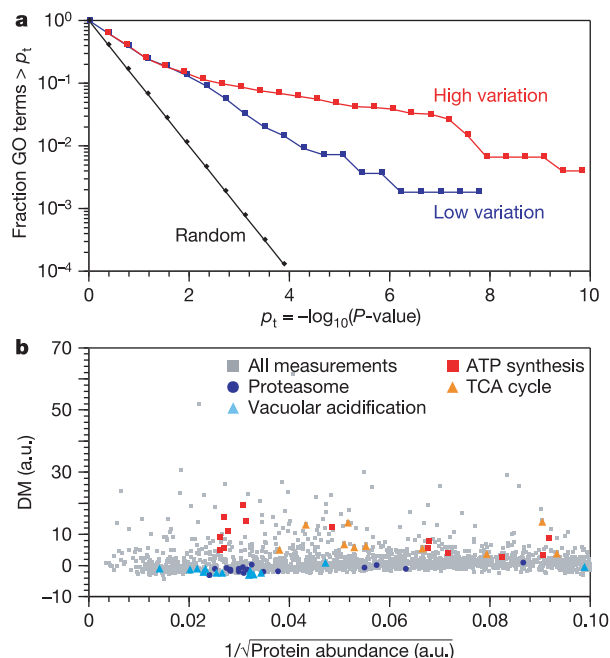


Figure 3 | Biological structure of protein variation. **a**, Biological processes (represented by Gene Ontology (GO) terms) are strongly enriched for proteins with low or high variation (blue and red, respectively). The distribution of P -values obtained using randomized data is shown in black (see also Supplementary Notes 2). **b**, Specific biological processes have higher or lower variation than expected by chance. Two classes of proteins with low noise participate in protein destruction. Two classes of proteins with high noise are involved in energy production.

influence the noise levels of individual proteins have only mild statistical correlations with low or high variation. These include chromosomal proximity of genes²² ($P \approx 0.21\text{--}3 \times 10^{-3}$; see Supplementary Notes 2), mRNA half-life ($P \approx 10^{-3}$), and participation in protein–protein interactions³⁷ ($P \approx 0.5$). Conversely, there is a dramatic enrichment of Gene Ontology (GO) terms³⁸ in both quiet and noisy proteins (Fig. 3a; Supplementary Table S5).

One of the most prominent correlations we find involves modes of transcriptional regulation and noise. For example, the binding of the transcription factors Fhl1, Rap1 and Abf1³⁹ to promoter regions is associated with low noise of the proteins encoded by the resulting gene products. In sharp contrast, numerous factors that act on chromatin structure to reversibly convert inactive DNA to active

DNA, including SAGA, SWI/SNF, Ino80, Isw2 and Swr1, regulate genes encoding proteins whose levels fluctuate considerably ($P = 10^{-36}$, $10^{-9.8}$, $10^{-6.7}$, $10^{-5.4}$ and $10^{-4.1}$, respectively). As suggested by Blake *et al.*²³ and Raser *et al.*³⁵, high noise is likely to be due, at least in part, to the introduction of a slow step into the production of mRNA, making the process more prone to bursts. In this respect, it is interesting to note that both Rap1 and Abf1 have been shown to disrupt nucleosome structure, possibly obviating the influence of chromatin in augmenting uncorrelated stochastic noise⁴⁰. Note, however, there are at least 80 proteins encoded by SAGA-dependent genes⁴¹ that have low noise ($DM < 1$), and this set is enriched for functional groups (for example, enzymes involved in glycolysis ($P \approx 10^{-10}$)). Conversely, there are at least 90 proteins

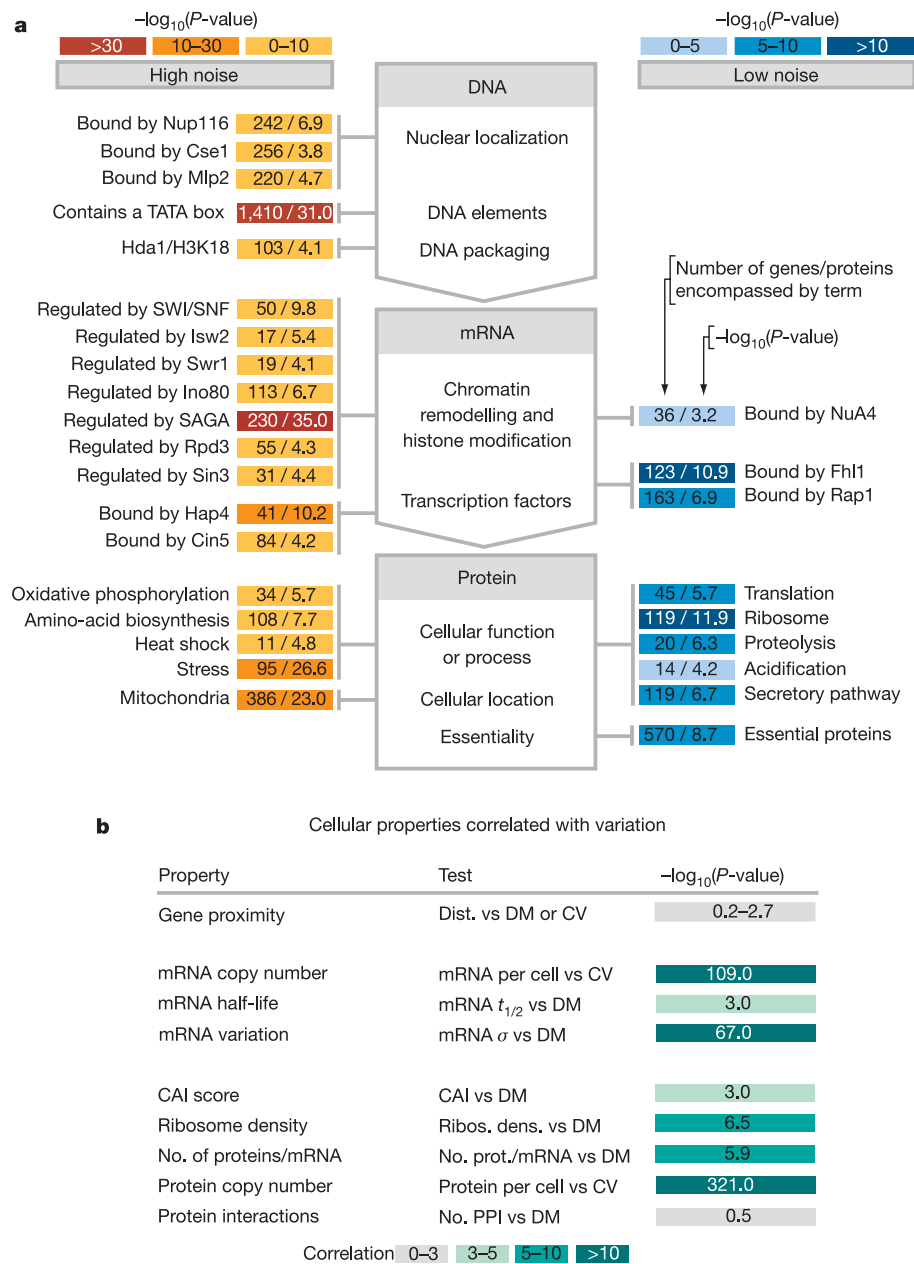


Figure 4 | Overview of major factors contributing to biological noise. **a**, Proteins targeted by chromatin remodelling complexes exhibit large variation whereas proteins participating in translation exhibit low variation (see also Supplementary Notes 2, Supplementary Fig. S14, Supplementary Table S5 and Supplementary Table S6). **b**, mRNA or protein copy number and the variation in mRNA expression are very strongly correlated with CV

and DM values, respectively. Other cellular properties previously postulated to influence noise^{22,37} are not. Abbreviations: Dist., distance; CV, coefficient of variation; $t_{1/2}$, half-life; σ , standard deviation; DM, distance to median (see the main text for a definition); CAI, codon adaptation index; Ribos. dens., ribosome density; prot., protein; No., number; PPI, protein–protein interactions; vs, versus.

with high noise ($DM > 1$), including enzymes that participate in ATP synthesis ($P \approx 10^{-5}$), whose corresponding genes are not known to be targets of any of the chromatin remodelling/modifying complexes listed above. Therefore chromatin remodelling/modifying complexes are neither sufficient nor necessary to generate noise *in vivo*.

Proteins belonging to defined functional groups also are distinguished by their noise levels (Fig. 3b and Fig. 4). For example, the SAGA-dependent genes encoding proteins that respond to changes in the environment, such as those involved in stress-response, amino acid biosynthesis, and in heat shock ($P \approx 10^{-27}$, $P \approx 10^{-8}$ and $P \approx 10^{-5}$, respectively), are correlated with high levels of noise. Conversely, proteins involved in translation initiation ($P \approx 10^{-5}$), translation/ribosomal proteins ($P \approx 2 \times 10^{-6}$) and protein degradation ($P \approx 5 \times 10^{-7}$) exhibit low variation. In addition to disruption of nucleosome structure (see above), the low noise in ribosomal proteins also might arise from tight coupling of ribosome number to growth status and cell size, and/or the rapid degradation of free ribosomal subunits (see ref. 42).

Finally, we find an unexpected relationship between noise and cellular location. Proteins localized to the Golgi (and, in particular, COPI vesicles) exhibit low variation ($P \approx 10^{-5}$). In contrast, high variation is correlated with localization to some membrane-bound organelles such as mitochondria ($P \approx 10^{-23}$) and peroxisomes ($P \approx 10^{-3}$) (Supplementary Table S5), raising the possibility that noise can be caused by unequal partitioning of low-copy number organelles during mitosis. Further evidence of this is provided by our observation that strains expressing GFP fused to an intact peroxisomal targeting sequence (GFP-PTS1) have higher noise than those expressing GFP fused to a reversed sequence comprising the same amino acids (GFP-pts1*), even when the proteins are expressed from identical promoters and loci (Supplementary Notes 2 and Supplementary Fig. S11a). Additionally, noise in peroxisomal proteins is specifically increased by disruption of correct peroxisome partitioning (by deletion of *Inp1*⁴³; Supplementary Fig. S11b). Thus, protein localization—independent of abundance or mode of transcription—is sufficient to influence single-cell variation.

Discussion

Here we present a new strategy for large-scale protein abundance measurements that is reproducible, sensitive, rapid and simple. The immediate value of our bulk protein measurements is demonstrated by their ability to characterize the response of cells grown in different environments and to provide a substantially richer view of cellular behaviours than could have been obtained from monitoring mRNA levels alone. The examples of post-transcriptional regulation identified here undoubtedly represent only a small fraction of the cases where mRNA and protein levels are discordant, as steady-state growth is likely to favour correlated transcription and translation. Indeed, we observe that under dynamic conditions, mRNA and protein levels are often markedly different (J.R.S.N., S.G. and J.S.W., unpublished data).

Perhaps the most fundamental advance reported here is the analysis of thousands of strains at single-cell resolution, permitting us to define the principal sources of protein noise. These studies reveal that the major factor governing protein variation is abundance. Moreover, the shape and magnitude of the global relationship between noise and abundance strongly argues that variation most likely originates from the stochastic production and destruction of mRNA molecules. Indeed, the magnitude of the variation observed here ($CV \sim 30\%$ for low–medium abundant proteins) is entirely consistent with that expected if protein variation results from Poisson noise owing to small mRNA numbers (1–2 per cell⁴⁴) and is mitigated by a filtering effect that arises because proteins are typically far longer-lived than their messages⁴⁵ (see Supplementary Notes 2).

Importantly, this concentration-dependent noise represents a

fundamental limitation on expression precision. For mechanisms involving the concerted production of messages (for example, following chromatin remodelling; see below), noise levels might increase above this limit. Conversely, if there were a biological need for lower noise, this could only be achieved through a specific mechanism (for example, feedback control or very short-lived messages)²⁶. Consistent with this interpretation, there is an effective lower bound for variation ($DM \sim -4$), but a much broader range of DM values is associated with high variation.

Beyond such global features of noise described above, our studies reveal a remarkable structure in protein-specific variation, and point to a central role for transcriptional regulation in determining this structure. Thus, noise, in addition to other considerations such as speed, frequency and amplitude, is a critical property that must be described when defining the response inherent to a given transcriptional mechanism.

In this context, it is revealing to look at the mechanisms of transcriptional regulation used by different functional classes of proteins. The use of a mode of transcription that helps minimize noise for polypeptides governing protein production and destruction (for example, potentially through the disruption of chromatin structure) may guarantee that downstream processes that must be effected accurately (for example, the cell cycle) are not burdened by imprecision³⁶. In contrast, proteins that change abundance with changing environmental conditions are controlled preferentially by factors begetting large amounts of variation. This noise might be an acceptable by-product of the mode of transcription: a large dynamic range, achieved through the ability to undergo strong repression or activation, is obtained at the cost of large variation. In fact, genes with the largest dynamic ranges encode proteins with the greatest noise ($P \approx 10^{-67}$; Supplementary Fig. S15). However, a large dynamic range can be achieved without a concomitant increase in noise (for example, ribosomal proteins). This suggests an alternative explanation: for some proteins that permit cells to respond to environmental perturbations, excursions from the mean at the single-cell level might benefit populations. In the short term, such deviations might facilitate a cell's initial response to environmental variation. More generally, the capacity to vary might permit a population to sample multiple phenotypic states to maximize the chances of some, but not all, cells' survival in an adverse environment¹³.

In summary, we have outlined and validated an approach that makes the yeast proteome readily accessible to quantitative single-cell measurements. As new GFP libraries become available, it should be straightforward to extend our approach to other organisms. Additionally, we have provided the first overarching view of noise in *S. cerevisiae*. The establishment of a robust strategy to rapidly measure both intrinsic and extrinsic noise with high precision across many strains now puts us in an excellent position to broadly explore the biology of noise.

METHODS

Strains. Strain acquisition, construction, growth, handling and genotypes are described in Supplementary Information.

Flow cytometry. Cells were delivered to an analytical cytometer using an autosampler device (LSR-II and HTS, respectively; Becton Dickinson). The autosampler was controlled by custom software (HTS-Pro; see Supplementary Notes 1 and 2). GFP was excited at 488 nm, and fluorescence was collected through a 505-nm long-pass filter and either a HQ510/20X or a HQ515/20X band-pass filter (Chroma Technology Corp.). tdTomato²⁹ was excited at 532 nm, and fluorescence was collected through a 585/42-nm band-pass filter.

Data analysis. Raw cytometry data were processed to eliminate systematic errors due to (1) uneven sample flow, (2) events at the bottom or top of the instrument's range, and (3) rare events unlikely to represent normal cellular properties (as judged either by size or fluorescence). Processed data were used for bulk fluorescence calculations. Calculated values were rejected if additional quality control thresholds were not met (for example, repeat measurements did not agree to within $\pm 20\%$; see Supplementary Information). Single-cell variation was calculated for cells contained within a circle defined on the FSC/SSC

plane having a radius of 4,096 arbitrary units (a.u.) and centred about the medians of these two parameters. Intrinsic and extrinsic noise values were calculated as described by Elowitz *et al.*²⁴ using strains containing GFP- and tdTomato-tagged proteins.

Additional techniques. Additional experimental procedures are listed in Supplementary Information.

Received 6 December 2005; accepted 6 April 2006.

Published online 14 May 2006.

- Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
- Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
- Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
- Andersen, J. S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
- Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
- Wei, J. *et al.* Global proteome discovery using an online three-dimensional LC–MS/MS. *J. Proteome Res.* **4**, 801–808 (2005).
- Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
- Gygi, S. P., Rochon, Y., Franz, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
- Washburn, M. P. *et al.* Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **100**, 3107–3112 (2003).
- Schrödinger, E. *What is Life? The Physical Aspect of the Living Cell* (Cambridge Univ. Press, Cambridge, UK, 1944).
- Barkai, N. & Leibler, S. Circadian clocks limited by noise. *Nature* **403**, 267–268 (2000).
- Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science* **305**, 1622–1625 (2004).
- Samoilov, M., Pilyasunov, S. & Arkin, A. P. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc. Natl Acad. Sci. USA* **102**, 2310–2315 (2005).
- Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).
- Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc. Natl Acad. Sci. USA* **43**, 553–566 (1957).
- Ferrell, J. E. Jr & Machleder, E. M. The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* **280**, 895–898 (1998).
- Biggar, S. R. & Crabtree, G. R. Cell signaling can direct either binary or graded transcriptional responses. *EMBO J.* **20**, 3167–3176 (2001).
- Lahav, G. *et al.* Dynamics of the p53–Mdm2 feedback loop in individual cells. *Nature Genet.* **36**, 147–150 (2004).
- Eitzman, P. D., Hendrick, J. L. & Srien, F. Quantitative immunofluorescence in single *Saccharomyces cerevisiae* cells. *Cytometry* **10**, 475–483 (1989).
- Edwards, B. S., Kuckuck, F. & Sklar, L. A. Plug flow cytometry: an automated coupling device for rapid sequential flow cytometric sample analysis. *Cytometry* **37**, 156–159 (1999).
- Becskei, A., Kaufmann, B. B. & van Oudenaarden, A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genet.* **37**, 937–944 (2005).
- Blake, W. J., Kaern, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
- Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
- Paulsson, J. Prime movers of noisy gene expression. *Nature Genet.* **37**, 925–926 (2005).
- Zaslaver, A. *et al.* Just-in-time transcription program in metabolic pathways. *Nature Genet.* **36**, 486–491 (2004).
- Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).
- Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotechnol.* **22**, 1567–1572 (2004).
- Wu, J. Q. & Pollard, T. D. Counting cytokinesis proteins globally and locally in fission yeast. *Science* **310**, 310–314 (2005).
- Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Wodicka, L., Dong, H., Mittmann, M., Ho, M. H. & Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
- Felice, M. R. *et al.* Post-transcriptional regulation of the yeast high affinity iron transport system. *J. Biol. Chem.* **280**, 22181–22190 (2005).
- Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**, 699–706 (2005).
- Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
- Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
- Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. Noise minimization in eukaryotic gene expression. *PLoS Biol.* **2**, e137 (2004).
- Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
- Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
- Yu, L. & Morse, R. H. Chromatin opening and transactivator potentiation by RAP1 in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19**, 5279–5288 (1999).
- Huisinga, K. L. & Pugh, B. F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in *Saccharomyces cerevisiae*. *Mol. Cell* **13**, 573–585 (2004).
- Warner, J. R. Synthesis of ribosomes in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **53**, 256–271 (1989).
- Fagarasanu, M., Fagarasanu, A., Tam, Y. Y., Aitchison, J. D. & Rachubinski, R. A. Imp1p is a peroxisomal membrane protein required for peroxisome inheritance in *Saccharomyces cerevisiae*. *J. Cell Biol.* **169**, 765–775 (2005).
- Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
- Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA* **99**, 5860–5865 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors acknowledge M. Bigos, P. Dezain and S. Elmes for their help with cytometry; K. Uffenheimer and A. Carroll (A.C.) for assistance with automation; W. Wickner for an anti-GFP antibody; R. Tsien for a construct encoding tdTomato; A.C., W.-K. Huh, M. Jonikas, V. Zapeda and E. Griffis for experimental assistance; A. H. DePace for graphical assistance; S. Collins, V. Denic, H. El Samad, V. L. Newman, E. K. O'Shea and members of the Weissman laboratory for insightful comments; and the Hertz Foundation, the NIH, DARPA and HHMI for funding.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.R.S.N. (jrsnewman@alum.mit.edu) or J.S.W. (weissman@cmp.ucsf.edu).

Distribution and three-dimensional structure of AIDS virus envelope spikes

Ping Zhu¹, Jun Liu¹, Julian Bess Jr², Elena Chertova², Jeffrey D. Lifson², Henry Grisé¹, Gilad A. Ofek³, Kenneth A. Taylor¹ & Kenneth H. Roux¹

Envelope glycoprotein (Env) spikes on AIDS retroviruses initiate infection of host cells and are therefore targets for vaccine development. Though crystal structures for partial Env subunits are known, the structure and distribution of native Env spikes on virions is obscure. We applied cryoelectron microscopy tomography to define ultrastructural details of spikes. Virions of wild-type human immunodeficiency virus 1 (HIV-1) and a mutant simian immunodeficiency virus (SIV) had ~ 14 and ~ 73 spikes per particle, respectively, with some clustering of HIV-1 spikes. Three-dimensional averaging showed that the surface glycoprotein (gp120) 'head' of each subunit of the trimeric SIV spike contains a primary mass, with two secondary lobes. The transmembrane glycoprotein 'stalk' of each trimer is composed of three independent legs that project obliquely from the trimer head, tripod-like. Reconciling available atomic structures with the three-dimensional whole spike density map yields insights into the orientation of Env spike structural elements and possible structural bases of their functions.

The Env spikes of HIV-1 are composed of surface envelope glycoprotein (gp120^{SU}) and transmembrane envelope glycoprotein (gp41TM) subunits, which, in their native configuration, elicit partially effective humoral immune responses^{1,2} and represent obvious vaccine targets for viral prophylactics^{3–6}. The Env spikes are thought to be trimeric^{7–10} and structure-based models have been proposed^{11–13}. However, despite intensive efforts, the arrangement and orientation of the loop-deleted gp120 core atomic structures within a native Env spike and their association with the gp41 subunits have remained largely speculative^{12,13}. We recently obtained electron tomograms of the Env spikes on intact, chemically fixed SIV and HIV-1, preserved in negative stain, which clearly reveal trimeric spikes on the viral surface¹⁴. Because of the staining, fixation and drying artefacts that can be associated with this method, some potentially important features of the virions and their associated trimers could not be deduced. We have now applied cryo-electron microscopy (cryoEM) tomography¹⁵ to investigate the surface features of unfixed, unstained frozen hydrated HIV-1 and SIV virions to reveal the distribution pattern of Env spikes on individual virions and to generate a three-dimensional (3D) model of the SIV Env spike. We have also attempted to reconcile this model of authentic virion trimers with available atomic structures of their constituent subunits.

Spike numbers and distribution

The cryoEM tomograms show that virions in our SIV and HIV-1 samples were intact and abundant with diameters of 109 ± 14 and 110 ± 8 nm, respectively (Fig. 1 and, for SIV, Supplementary Movie 1). Importantly, Env spikes are readily seen on virion surfaces (arrowheads). The SIV mac239 variant virus studied here has a mutation resulting in a transmembrane glycoprotein with a truncated cytoplasmic tail resulting in the elimination of endocytosis signals and a high rate of cell surface expression and viral incorporation^{16,17}. CryoEM of the SIV virions (Fig. 1a, Fig. 2a and

Supplementary Movie 1) showed a surface fairly well saturated with 73 ± 25 ($n = 52$) rather uniformly distributed Env spikes, in line with our previous data¹⁴. A nearest-neighbour analysis of interspike centre-to-centre spacing showed a distribution with a

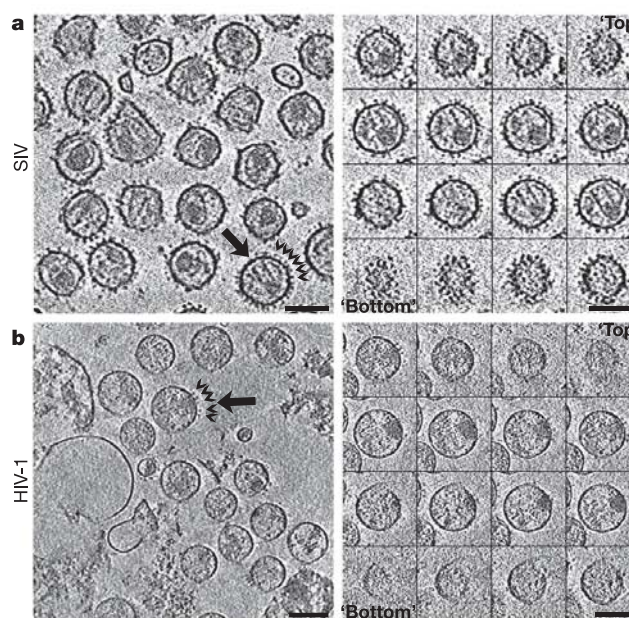


Figure 1 | Representative tomographic images of mutant SIV and wild-type HIV-1. Shown are computationally derived transverse sections through fields of SIV (**a**, left) and HIV-1 (**b**, left) virions and sequential Z-stack tomogram layers of selected virions (arrows on the left): **a** right, SIV Z-stack; **b** right, HIV-1 Z-stack. Examples of putative Env spikes on selected virions are indicated by arrowheads on the left. Scale bars, 100 nm.

¹Department of Biological Science and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306, USA. ²AIDS Vaccine Program, SAIC Frederick Inc, National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. ³Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.

peak centring on ~ 15 nm (Supplementary Fig. S1a). No evidence of discrete periodic peak distances was observed (Supplementary Fig. S1d) as might have been expected if spike placement reflected an underlying geometry of matrix/capsid structure.

In contrast to SIV, yet consistent with our earlier results¹⁴, cryoEM analysis of wild-type HIV-1 virions ($n = 40$), which have a native full-length gp41TM cytoplasmic tail, showed only 14 ± 7 Env spikes per particle (range 4 to 35) (see examples in Fig. 2b–d). As with SIV, no clear evidence of periodic spacing was observed (Supplementary Fig. S1e and f, respectively). The most frequent nearest-neighbour interspike spacing was ~ 15 nm (Supplementary Fig. S1b) which was closer than was predicted (~ 23 nm) from a virion model with randomly distributed spikes (see Supplementary Fig. S1c). A cluster analysis (Supplementary Fig. S2) also showed a higher proportion of closely spaced spikes than was predicted from chance alone (see examples in Fig. 2b and c). For example, 8.2% and 16.5% of the observed spikes were in clusters of at least seven members linked by distances of ≤ 22.5 nm and ≤ 30 nm, respectively. In contrast, 0.6% and 4.8% of the spikes were similarly linked in the random distribution model (Supplementary Fig. S2).

Greater numbers of spikes per virion are correlated with increased infectivity¹⁸. Fusion efficiency might also be dependent upon local spike density (clustering), especially if concerted receptor engagement of multiple Env spikes is required for binding and fusion^{10,19,20}. Spike clustering may also facilitate antibody (Ab)-mediated neutralization by fostering bivalent cross-linking^{21–23}. A paratope-to-paratope reach of 10–20 nm for typical Abs is well within the range of at least some epitopes on spikes 15 nm apart.

The clustering data are consistent with the hypothesis that Env spikes are not free to diffuse in the plane of the membrane. However, the lack of both an apparent ordered array of spikes and of evidence of periodicity in the inter-spike distances conflicts with the notion that the tails are inserted into a geometrically arrayed set of matrix pores^{24,25}. Clustering need not reflect matrix association but could suggest a direct inter-spike interaction mediated by the cytoplasmic tails themselves.

The Env spike tomogram average and a spike model

We focused our initial cryoEM tomography efforts on the SIV virions with truncated transmembrane glycoproteins, as the significantly higher numbers of spikes per virion facilitated the analysis¹⁴. Though it seems unlikely, we cannot rule out the possibility that the lack of a full-length cytoplasmic tail has, in some way, influenced our cryoEM Env spike structure or that there could be significant structural differences between SIV and HIV-1 spikes.

To generate a model of the Env spike, 6,175 individual SIV spike

volumes from 122 virions were aligned and averaged. On the basis of our earlier studies¹⁴ and the inherent three-fold symmetry in class averages of our cryoEM data obtained by correspondence analysis (Supplementary Methods and Fig. S3), the spike volume average was three-fold symmetrized to generate the final averaged trimer density map (Supplementary Fig. S4). Thus, each monomeric subunit within the final density map represents the data from 18,525 (that is, $6,175 \times 3$) individual monomeric subunits with a measured resolution of ~ 3.2 nm. Eight unsymmetrized class averages (Supplementary Fig. S3) were not remarkably different from the three-fold symmetrized global average (Supplementary Fig. S4), suggesting a relatively homogeneous spike population at the resolution achieved.

As shown in Fig. 3a–c and Supplementary Movie 2, each Env spike subunit contains three lobes, designated ‘main’, ‘lateral’ and ‘proximal’. Projecting from the three main lobes of the spike are ridges (‘arches’ in Fig. 3d) projecting medio-distally which join to form an apical ‘peak’ on the three-fold axis. The apical peak obscures an axial region of lower density; probably representing a hollow cavity just below it, a feature more apparent in a surface generated using a higher density threshold (Fig. 3d). Emanating from each proximal lobe is a short ‘leg’ that angles outward, whereupon it meets the viral ‘membrane’ and continues laterally just above the plane of the membrane to form a ‘foot’. The three splayed legs of the spike create a distinct cavity at the three-fold axis just above the membrane (Fig. 3a and e).

The height (13.7 nm) of the spike, the diameter of the head (10.5 nm), and the vertical height of the stalk (1.9 nm) are all in line with our previous measurements and most of those in the

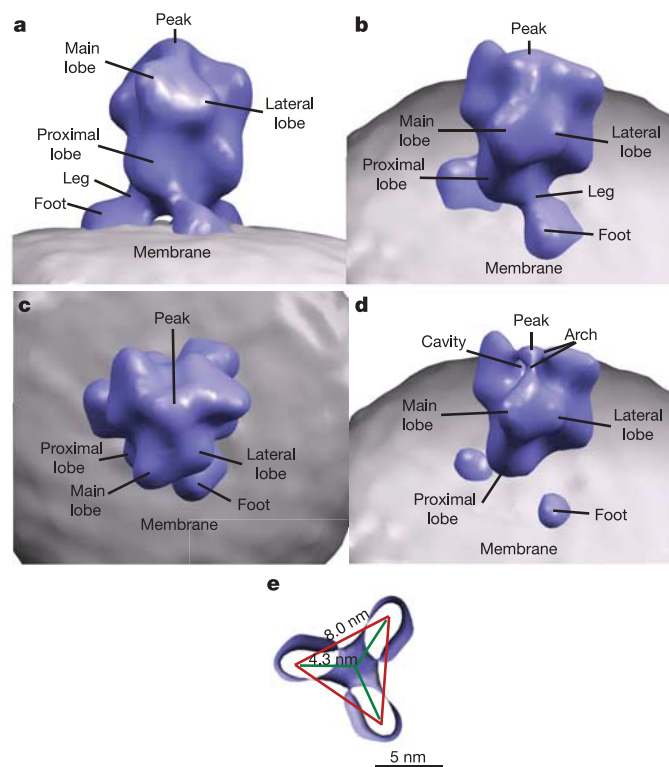


Figure 3 | Surface-rendered model of averaged SIV Env spike tomograms.

Env spike (blue) and membrane (grey) in side (a), intermediate (b), and top (c) views and an intermediate view at a higher density threshold (d). Panel e represents a membrane-parallel slice encompassing the legs and feet (MPER) as viewed from below. The centre-to-centre distance between possible MSD attachment sites (red lines) and the medial axis to MSD sites (green lines) are indicated. The spike and membrane surfaces (in a–d) were rendered at different thresholds, the latter adjusted to conform to the known thickness of the lipid bilayer (~ 4 nm).

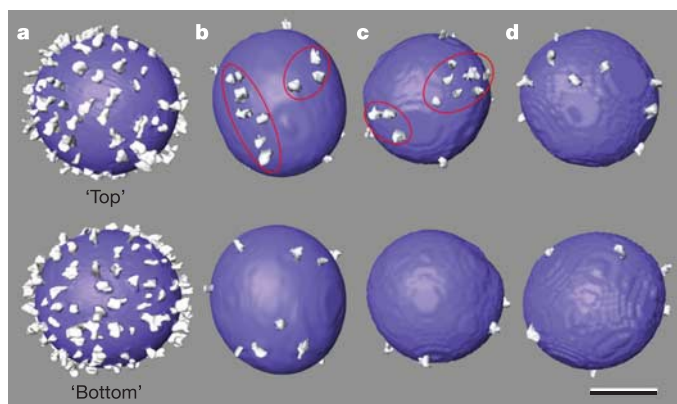


Figure 2 | Surface-rendered models of a representative SIV (a) and three HIV-1 (b–d) virions with highlighted (white) presumptive Env spikes.

The two hemispheres of each virion are arbitrarily labelled as ‘Top’ and ‘Bottom’. Ellipses encircle clusters of three or more spikes. Scale bar, 50 nm.

literature^{3,14}. The diameter of each individual leg, a feature not previously described, ranges from about 1.9 to 3.7 nm. Because this region of the spike is believed to be rather flexible⁸, the features of the feet and legs in the model could be influenced by local density distortions resulting from averaging of these features when in different orientations.

The Moloney murine leukaemia virus (MMLV), the only other retroviral Env spike similarly analysed to date²⁶, also shows three-fold symmetry. The MMLV Env spike appears to be squatter ($\sim 10 \times 10$ nm) than the SIV spike—with multiple, more pronounced, distal and lateral projections²⁶. Interestingly, the Env spikes on both viruses display an open tripod-like leg configuration. Thus, as discussed below, it appears that the membrane proximal regions of the three transmembrane glycoproteins of each trimer are not intimately associated in parallel in a combined stalk orthogonal to the viral envelope as is generally portrayed^{5,27}.

The gp120 triad

There are now two 3D models of the gp120 trimer based on the crystal structures of variable loop-deleted AIDS virus gp120 core fragments: SIV in the pre-liganded presumptive native state conformation¹³, and HIV-1, liganded with CD4 and an anti-gp120 Fab fragment^{11,12}. SIV and HIV-1 Env are considered similar enough in sequence to allow valid structural comparisons¹³. However, the trimer models vary considerably from one another, based largely on structural differences associated with large conformational changes induced by receptor binding. Consequently, only the unliganded SIV model may be appropriately compared to our unliganded SIV cryoEM density map.

An initial attempt to fit the unliganded SIV gp120 core trimer model (missing structures including the V1/V2 and V3 loops) of Chen *et al.* (figure 6 in ref. 13) into our averaged cryoEM Env spike density resulted in a considerable proportion of the atomic structure protruding from the periphery (Supplementary Fig. S5a). Efforts to

move the protomers medially improved the fit somewhat but appeared sterically to restrict access to the disto-medially oriented CD4 binding sites (Supplementary Fig. S5b).

In a second fitting exercise (Fig. 4 and Supplementary Movie 3) we started afresh, attempting to optimize the positioning of the gp120 core structures (Fig. 4a, b, d–f) within the main lobes of the density map while maintaining orientation of the truncated amino and carboxy termini and the unstructured segment of the C2 loop downward and towards the central three-fold axis, where contact with gp41 presumably occurs¹³. This positioning minimized the loops and side chains protruding from our cryoEM envelope (Fig. 4a, d and Supplementary Fig. S5c) and positioned the CD4 binding sites on the periphery of the spike. The optimum fit aligned the truncated stems of the V1/V2 and V3 loops towards the otherwise unoccupied proximal and lateral lobes, respectively (Fig. 4a–e). The apical peak is shown unoccupied by the atomic model, consistent with its low density (Fig. 4f). The arch in Fig. 3d may represent an alternative position of the flexible V3 loop (that is, projecting towards the target cell surface) on a minority of the Env subunits, a conformation recently defined at the atomic level for the V3 loop in the CD4-bound fusion-active HIV-1 protomer²⁸.

The transmembrane glycoprotein

The fusion-active conformation of the transmembrane glycoprotein trimer is believed to be a rod-like extended coiled coil dominated by two α -helical regions, the N-terminal heptad repeat 1 (HR1) and C-terminal heptad repeat 2 (HR2) within each subunit. In contrast, the post-fusion conformation is defined by the 6-helix bundle wherein the HR2 regions have folded back on the HR1 regions, presumably destabilizing and drawing the viral and cell membranes together in the process^{7–9,29}. The prefusion structure of the transmembrane glycoprotein is considerably less well defined³⁰. Ab binding studies demonstrate that the region encompassed by HR1 and HR2 is essentially obscured in the native Env spike and so is not likely

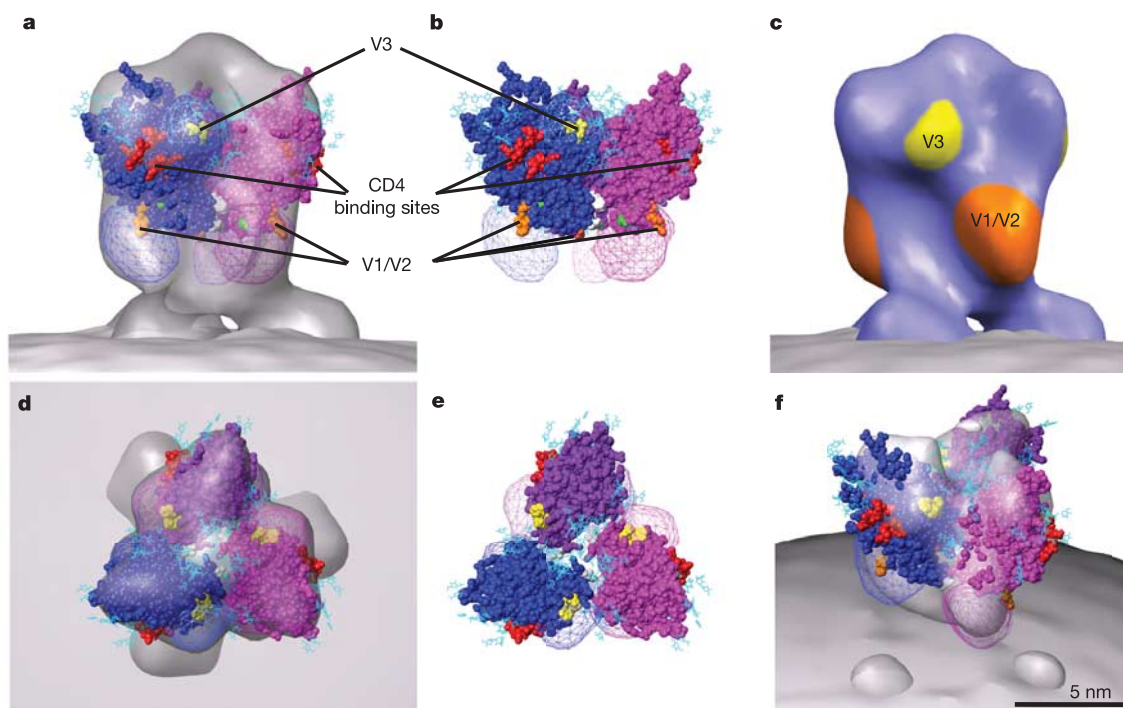


Figure 4 | Surface-rendered model of SIV Env spike manually fitted with SIV unliganded core atomic model. Atomic model: Protein Data Bank ID: 2BF1 (ref. 13). Side (a, b) and top (d, e) views, with (a, d) and without (b, e) enshrouding cryoEM surface. Indicated structures: gp120 subunits (blue, magenta, mauve), CD4 binding sites (red), V3 loop stems (yellow), V1/V2

loop stems (orange), N and C termini (green), C2 loop stems (white), proposed V3 and V1/V2 loop sites (mesh-enclosed volumes over stems or (in c) coloured surfaces). f, Spike as in a, but tilted and at higher density threshold, showing low-density region under peak.

to contribute directly to the surface features of the Env spike³⁰. This then leaves the membrane proximal external region (MPER: the segment between HR2 and the membrane-spanning domain, MSD) as a potential target for neutralizing Ab and the probable contributor to the bulk of the Env stalk observed by cryoEM.

The MPER is the most highly conserved region in the transmembrane glycoprotein²⁷ and has been proposed to serve as (1) a flexible extender⁸, (2) a contributor to trimer formation and stability^{31,32}, and (3) an agent for membrane destabilization that fosters membrane fusion²⁷. The amphipathic nature of the C-terminal portions of the MPER and the tendency for peptides of this region to self-associate have been interpreted as indicating that this region makes an important contribution to Env spike trimerization³¹. Our data contradict this notion. The lower density that we observe between the legs of the stem region, which appears as a pronounced gap in the surface-rendered model (Fig. 3a, e) indicates that this region of the transmembrane glycoprotein subunits is not self-associated in native trimers on prefusion virions but rather is engaged in an extensive interaction with the plasma membrane. In support of this interpretation are the observations that the anti-MPER monoclonal Abs (MAbs) 2F5 and 4E10 bind better to their respective epitopes when the epitopes are associated with lipid^{33,34}. Moreover, the 2F5 and 4E10 epitopes and, remarkably, the tips of the 2F5 and 4E10 MAbs themselves, display pronounced hydrophobic patches^{34,35}.

We manually mapped the 2F5 and 4E10 peptides into our SIV spike model under the assumption that this highly conserved region would be similarly configured in HIV-1 and SIV. In support of this assumption, both epitopes, when grafted into SIV, have been shown to be immunologically indistinguishable from those in HIV-1 (ref. 36). The placement of the 2F5 and 4E10 peptides in our map was facilitated by positioning the relevant hydrophobic residues of the peptides and their respective MAbs in the plane of the membrane. We found that the splayed legs of the tomogram model readily accommodated the peptides and are consistent with the cognate MAbs pointing towards the membrane (rather than towards a compact stalk) (Fig. 5 and Supplementary Movie 4).

The hydrophobicity of the MPER is a highly conserved feature of SIV, MMLV, HIV, and other retroviruses. Both SIV and MMLV, the two retroviruses for which cryoEM data are available, display splayed legs wherein the proximal regions seem to be associated with the membrane. Within HIV-1, the 4E10 epitope may be universally expressed³⁷, implying a strong conservation of function for this region. Together, these observations suggest a critical role for an association between the MPER and the outer leaflet of the viral membrane in viral function.

The MSD and matrix association

The MSD is generally believed to be an α -helical region composed

primarily of neutral and hydrophobic amino acids³⁸ which serves to stabilize the trimer by self-association. However, steric considerations, as outlined above, suggest that the MSDs are most probably displaced a considerable distance (~ 4.3 nm) from the three-fold axis (Figs 3a, e and 5a) with a likely centre-to-centre distance of ~ 8 nm. Consequently, any contribution to trimer stabilization in the prefusion configuration is probably due to simple membrane anchorage. It has been suggested that triggering of a single retrovirus spike is sufficient to initiate membrane fusion³⁹, in contrast to the engagement of multiple spikes as, for example, is believed necessary for influenza virus fusion^{40,41}. If so, then perhaps the three splayed legs with their unassociated MSDs, in conjunction with membrane-associated elements of the MPERs, provide the added degree of lipid anchorage required by some aspect of the fusion mechanism. One possible mechanism would be the fostering of more efficient and extensive recruitment and disruption of the outer leaflet of the viral membrane by the three widely spaced MPER/MSD structures of a single spike during the large conformational change in the Env spike that occurs upon triggering. In addition to mechanistic considerations, the open leg configuration could have important implications for the designs of future soluble Env trimer vaccine candidates, particularly those stabilized by engineered trimerization motifs.

Though we do not see evidence of it in mature virions, it is believed that the viral matrix proteins both interact with the cytoplasmic tails of Env and are organized into a regular geometric array^{24,42}. Nevertheless, the completeness of the array has been questioned. Models of matrix protein arrays, in which the resulting holes are spaced at about 7 nm intervals, have been proposed^{24,43}. Such patterns could accommodate 8 nm centre-to-centre spacing for MSDs if the cytoplasmic tails are angled slightly towards the three-fold axis.

Finally, it must be noted that only a minority of HIV-1 or SIV virions in typical preparations possess culturable infectivity, and it has been proposed that there may be non-functional forms of Env present on the virion surface⁴⁴. It is thus possible that the trimers of infectious particles may differ structurally from the averaged structures described here, although there are several factors that contribute to the low particle infectivity ratio of HIV-1 and SIV, many of which are unrelated to the viral Env. In any case, the current results provide a new level of structural information about native trimers on virions, and guidance for further studies.

Discussion

The present studies begin to bridge the gap between atomic level structures derived from crystallographic studies of partial Env spike subunits and the resolution of conventional transmission electron microscopy, revealing 3D ultrastructural details of individual native trimers. Additional studies of virions with variable loop deletions and of virions complexed with Env-targeted ligands of structurally

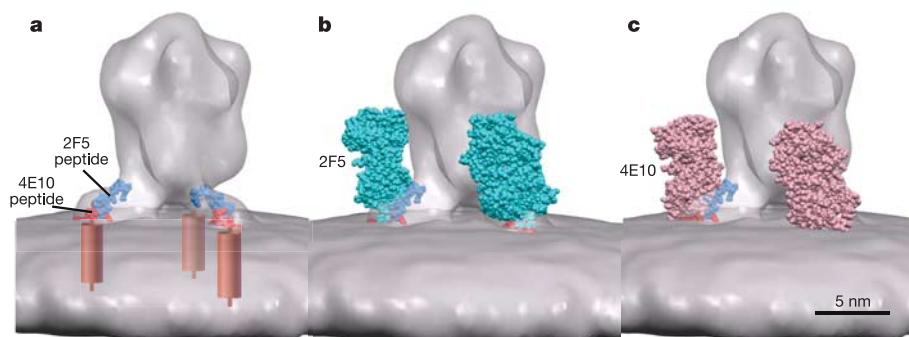


Figure 5 | Fitting of 2F5 and 4E10 peptides. Surface-rendered model of the SIV Env spike showing proposed placement of 2F5 (blue; PDB ID: 1TJ1) and 4E10 (red; PDB ID: 1TZG) HIV-1 epitope peptides in the leg and foot densities (a) based on alignment of hydrophobic residues in the peptides

and Abs parallel to the membrane. The proposed positions of the corresponding antigen-binding fragment (Fab) 2F5 (turquoise, b) and Fab 4E10 (pink, c), are shown. Red cylinders indicate proposed positions of the helical MSDs relative to the spike feet.

defined specificity should provide additional insights relevant to further understanding of virus binding and fusion and Ab neutralization.

METHODS

More detailed Methods are available as Supplementary Information.

Viruses. Highly purified virus samples⁴⁵ treated with Aldrichiol-2 (ref. 46) were provided by the AIDS Vaccine Program (SAIC Frederick, NCI, Maryland, USA).

CryoEM preparation. Virus aliquots were washed and placed on EM grids for 1 min. Excess virus and buffer were removed by blotting, and the sample rapidly vitrified. The grids were either examined immediately or stored in liquid nitrogen for later use.

Cryo-electron tomography. The EM grids were examined under low-dose conditions. Single axis tilt series of 70–80 images were recorded at 43,200X magnification and an underfocus of 4–6 µm.

Image analysis. Tomograms were computed by weighted back-projection using the PROALIGN program package⁴⁷ and imported into the software package AMIRA for visualization and surface rendering.

SIV spike volumes, manually selected on the basis of appearances, size and location, were subjected to 3D alignment and averaging algorithms⁴⁸. The morphologically most divergent 2% were excluded. The remaining spikes were aligned against a three-fold symmetrized reference and averaged. The average was then three-fold symmetrized to generate the final averaged trimer density map (Supplementary Fig. S4). The Fourier transform extended to 2.5 nm by Fourier Shell Correlation. The minimum distance measured between discrete internal density peaks was 3.2 nm (Supplementary Fig. S6).

Atomic model fitting. The indicated X-ray structures were manually fitted into the 3D averaged Env spike CryoEM density map using the Chimera Software package (<http://www.cgl.ucsf.edu/chimera/>).

Env spike distribution analyses. Neighbouring spikes within 22.5 nm or 30 nm were considered members of a cluster that included all other spikes similarly linked. The observed spike distribution data were compared with mathematically generated random distributions as described in Supplementary Fig. S1.

Received 8 March; accepted 24 April 2006.

Published online 24 May 2006.

- Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422**, 307–312 (2003).
- Richman, D. D., Wrinn, T., Little, S. J. & Petropoulos, C. J. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl Acad. Sci. USA* **100**, 4144–4149 (2003).
- Poignard, P., Saphire, E. O., Parren, P. & Burton, D. Gp120: biologic aspects of structural features. *Annu. Rev. Immunol.* **19**, 253–274 (2001).
- Parren, P. W. & Burton, D. R. The antiviral activity of antibodies *in vitro* and *in vivo*. *Adv. Immunol.* **77**, 195–262 (2001).
- Burton, D. R. *et al.* HIV vaccine design and the neutralizing antibody problem. *Nature Immunol.* **5**, 233–236 (2004).
- Mascola, J. R. *et al.* Protection of macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. *J. Virol.* **73**, 4009–4018 (1999).
- Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
- Weissenhorn, W., Dessen, A., Harrison, S. C., Skehel, J. J. & Wiley, D. C. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387**, 426–430 (1997).
- Caffrey, M. *et al.* Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *EMBO J.* **17**, 4572–4584 (1998).
- Doms, R. W. Beyond receptor expression: the influence of receptor conformation, density, and affinity in HIV-1 infection. *Virology* **276**, 229–237 (2000).
- Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**, 648–659 (1998).
- Kwong, P. D., Wyatt, R., Sattentau, Q. J., Sodroski, J. & Hendrickson, W. A. Oligomeric modeling and electrostatic analysis of the gp120 envelope glycoprotein of human immunodeficiency virus. *J. Virol.* **74**, 1961–1972 (2000).
- Chen, B. *et al.* Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* **433**, 834–841 (2005).
- Zhu, P. *et al.* Electron tomography analysis of envelope glycoprotein trimers on HIV and simian immunodeficiency virus virions. *Proc. Natl Acad. Sci. USA* **100**, 15812–15817 (2003).
- Lucic, V., Forster, F. & Baumeister, W. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.* **74**, 833–865 (2005).
- LaBranche, C. C. *et al.* A single acid change in the cytoplasmic domain of the simian immunodeficiency virus transmembrane molecule increases envelope glycoprotein expression on infected cells. *J. Virol.* **69**, 5217–5227 (1995).
- Kodama, T. *et al.* Significance of premature stop codons in *env* of simian immunodeficiency virus. *J. Virol.* **63**, 4709–4714 (1989).
- Yuste, E., Reeves, J. D., Doms, R. W. & Desrosiers, R. C. Modulation of Env content in virions of simian immunodeficiency virus: correlation with cell surface expression and virion infectivity. *J. Virol.* **78**, 6775–6785 (2004).
- Doms, R. W. & Moore, J. P. HIV-1 membrane fusion: targets of opportunity. *J. Cell Biol.* **151**, F9–F13 (2000).
- Kuhmann, S. E., Platt, E. J., Kozak, S. L. & Kabat, D. Cooperation of multiple CCR5 coreceptors is required for infections by human immunodeficiency virus type 1. *J. Virol.* **74**, 7005–7015 (2000).
- Cavacini, L. A., Emes, C. L., Power, J., Duval, M. & Posner, M. R. Effect of antibody valency on interaction with cell-surface expressed HIV-1 and viral neutralization. *J. Immunol.* **152**, 2538–2545 (1994).
- Klasse, P. J. & Sattentau, Q. J. Mechanisms of virus neutralization by antibody. *Curr. Top. Microbiol. Immunol.* **260**, 87–108 (2001).
- Yuste, E., Johnson, W., Pavlakis, G. N. & Desrosiers, R. C. Virion envelope content, infectivity, and neutralization sensitivity of simian immunodeficiency virus. *J. Virol.* **79**, 12455–12463 (2005).
- Forster, M. J., Mulloy, B. & Nermut, M. V. Molecular modelling study of HIV p17gag (MA) protein shell utilising data from electron microscopy and X-ray crystallography. *J. Mol. Biol.* **298**, 841–857 (2000).
- Hill, C. P., Worthylake, D., Bancroft, D. P. & Christensen, A. M. Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: Implications for membrane association and assembly. *Proc. Natl Acad. Sci. USA* **93**, 3099–3104 (1996).
- Förster, F., Medalia, O., Zauberman, N., Baumeister, W. & Fass, D. Retrovirus envelope protein complex structure *in situ* studied by cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **102**, 4729–4734 (2005).
- Zwick, M. B. *et al.* Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J. Virol.* **75**, 10892–10905 (2001).
- Huang, C. *et al.* Structure of a V3-containing HIV-1 gp120 core. *Science* **310**, 1025–1028 (2005).
- Tan, K., Liu, J.-H., Wang, J.-H., Shen, S. & Lu, M. Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proc. Natl Acad. Sci. USA* **94**, 12303–12308 (1997).
- Qiao, Z. *et al.* Design, expression, and immunogenicity of a soluble HIV trimeric envelope fragment adopting a prefusion gp41 configuration. *J. Biol. Chem.* **280**, 23138–23146 (2005).
- Salzwedel, K., West, J. T. & Hunter, E. A conserved tryptophan-rich motif in the membrane-proximal region of the human immunodeficiency virus type 1 gp41 ectodomain is important for Env-mediated fusion and virus infectivity. *J. Virol.* **73**, 2469–2480 (1999).
- Saez-Cirion, A. *et al.* Structural and functional roles of HIV-1 gp41 pretransmembrane sequence segmentation. *Biophys. J.* **85**, 3769–3780 (2003).
- Grundner, C., Mirzabekov, T., Sodroski, J. & Wyatt, R. Solid-phase proteoliposomes containing human immunodeficiency virus envelope glycoproteins. *J. Virol.* **76**, 3511–3521 (2002).
- Ofek, G. *et al.* Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. *J. Virol.* **78**, 10724–10737 (2004).
- Cardoso, R. M. *et al.* Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. *Immunity* **22**, 163–173 (2005).
- Yuste, E. *et al.* Simian immunodeficiency virus engrafted with human immunodeficiency virus type 1 (HIV-1)-specific epitopes: replication, neutralization, and survey of HIV-1-positive plasma. *J. Virol.* **80**, 3030–3041 (2006).
- Binley, J. M. *et al.* Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J. Virol.* **78**, 13232–13252 (2004).
- West, J. T., Johnston, P. B., Dubay, S. R. & Hunter, E. Mutations within the putative membrane-spanning domain of the simian immunodeficiency virus transmembrane glycoprotein define the minimal requirements for fusion, incorporation, and infectivity. *J. Virol.* **75**, 9601–9612 (2001).
- Yang, X., Kurteva, S., Ren, X., Lee, S. & Sodroski, J. Stoichiometry of envelope glycoprotein trimers in the entry of human immunodeficiency virus type 1. *J. Virol.* **79**, 12132–12147 (2005).
- Blumenthal, R., Sarkar, D. P., Durell, S., Howard, D. E. & Morris, S. J. Dilation of the influenza hemagglutinin fusion pore revealed by the kinetics of individual cell-cell fusion events. *J. Cell Biol.* **135**, 63–71 (1996).
- Danieli, T., Pelletier, S. L., Henis, Y. I. & White, J. M. Membrane fusion mediated by the influenza virus hemagglutinin requires the concerted action of at least three hemagglutinin trimers. *J. Cell Biol.* **133**, 559–569 (1996).
- Freed, E. O. & Martin, M. A. Domains of the human immunodeficiency virus type 1 matrix and gp41 cytoplasmic tail required for envelope incorporation into virions. *J. Virol.* **70**, 341–351 (1996).
- Rao, Z. *et al.* Crystal structure of SIV matrix antigen and implications for virus assembly. *Nature* **378**, 743–747 (1995).
- Moore, P. L. *et al.* Nature of nonfunctional envelope proteins on the surface of human immunodeficiency virus type 1. *J. Virol.* **80**, 2515–2528 (2006).

45. Chertova, E. *et al.* Sites, mechanism of action and lack of reversibility of primate lentivirus inactivation by preferential covalent modification of virion internal proteins. *Curr. Mol. Med.* **3**, 265–272 (2003).
46. Rossio, J. L. *et al.* Inactivation of human immunodeficiency virus type 1 infectivity with preservation of conformational and functional integrity of virion surface proteins. *J. Virol.* **72**, 7992–8001 (1998).
47. Winkler, H. & Taylor, K. A. Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography. *Ultramicroscopy* **106**, 240–254 (2006).
48. Winkler, H. & Taylor, K. A. Multivariate statistical analysis of three-dimensional cross-bridge motifs in insect flight muscle. *Ultramicroscopy* **77**, 141–152 (1999).

Supplementary Information is linked to online versions of the paper at www.nature.com/nature.

Acknowledgements We thank H. Winkler for assistance with volume alignment and classification and 3D volume pickup and P. D. Kwong for his comments on the manuscript. The work was supported in part by a NIH and NCI contract (J.B., E.C., J.D.L.) and grants from the NIH NIAID (P.Z., H.G., K.H.R.) and the NIH NIGMS (J.L., K.A.T.). This project has been funded in whole or in part with federal funds from the National Cancer Institute, NIH. The content of this

publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Author Contributions P.Z. performed the bulk of the cryoEM, image analysis, model fitting, and graphic illustration. J.L. assisted with the image acquisition, analysis and interpretation. J.B. cultured and purified the viruses. E.C. biochemically analysed the virus preparations for purity. J.D.L. contributed to the design of the project and contributed extensively to the manuscript writing. H.G. assisted with fitting of the 2F5 and 4E10 peptides and MAbs into the cryoEM density map. G.A.O. contributed to the analysis of the data, particularly as it relates to the transmembrane region of the envelope spike. K.A.T. provided technical expertise on various aspects of the cryoEM and tomography and contributed to technical portions of the manuscript. K.H.R. provided overall management of the project as well as data interpretation, manuscript writing and figure design.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.H.R. (roux@bio.fsu.edu).

Repulsively bound atom pairs in an optical lattice

K. Winkler¹, G. Thalhammer¹, F. Lang¹, R. Grimm^{1,3}, J. Hecker Denschlag¹, A. J. Daley^{2,3}, A. Kantian^{2,3}, H. P. Büchler^{2,3} & P. Zoller^{2,3}

Throughout physics, stable composite objects are usually formed by way of attractive forces, which allow the constituents to lower their energy by binding together. Repulsive forces separate particles in free space. However, in a structured environment such as a periodic potential and in the absence of dissipation, stable composite objects can exist even for repulsive interactions. Here we report the observation of such an exotic bound state, which comprises a pair of ultracold rubidium atoms in an optical lattice. Consistent with our theoretical analysis, these repulsively bound pairs exhibit long lifetimes, even under conditions when they collide with one another. Signatures of the pairs are also recognized in the characteristic momentum distribution and through spectroscopic measurements. There is no analogue in traditional condensed matter systems of such repulsively bound pairs, owing to the presence of strong decay channels. Our results exemplify the strong correspondence between the optical lattice physics of ultracold bosonic atoms and the Bose–Hubbard model^{1,2}—a link that is vital for future applications of these systems to the study of strongly correlated condensed matter and to quantum information.

Cold atoms loaded into a three-dimensional (3D) optical lattice provide a realization of a quantum lattice gas^{1,2}. An optical lattice can be generated by pairs of counterpropagating laser beams, where the resulting standing wave intensity pattern forms a periodic array of microtraps for the cold atoms, with period a given by half the wavelength of the light, $\lambda/2$. The periodicity of the potential gives rise to a band structure for the atom dynamics with Bloch bands separated by bandgaps, which can be controlled by the laser parameters and beam configuration. The dynamics of ultracold atoms loaded into the lowest band of a sufficiently deep optical lattice is well described by the Bose–Hubbard model with hamiltonian^{1,3}:

$$\hat{H} = -J \sum_{\langle i,j \rangle} \hat{b}_i^\dagger \hat{b}_j + \frac{U}{2} \sum_i \hat{n}_i (\hat{n}_i - 1) + \sum_i \varepsilon_i \hat{n}_i \quad (1)$$

Here $\hat{b}_i$ ($\hat{b}_i^\dagger$) are destruction (creation) operators for the bosonic atoms at site i , and $\hat{n}_i = \hat{b}_i^\dagger \hat{b}_i$ is the corresponding number operator. $J/\hbar$ denotes the nearest-neighbour tunnelling rate, U the on-site collisional energy shift, and ε_i the background potential. The high degree of control available over the parameters in this system—for example, changing the relative values of U and J by varying the lattice depth, V_0 —has led to seminal experiments on strongly correlated gases in optical lattices. These experiments include the study of the superfluid–Mott insulator transition⁴, the realization of one-dimensional (1D) quantum liquids with atomic gases^{5,6} (see also refs 7 and 8), and the investigation of disordered systems⁹. 3D optical lattices have also opened new avenues in cold collision physics and chemistry^{10–13}.

A striking prediction of the Bose–Hubbard hamiltonian (equation (1)) is the existence of stable repulsively bound atom pairs. These are most intuitively understood for strong repulsive interaction

$|U| \gg J$, $U > 0$, where an example of such a pair is a state of two atoms occupying a single site, $|2_i\rangle \equiv (\hat{b}_i^{\dagger 2} |\text{vac}\rangle)/\sqrt{2}$, where $|\text{vac}\rangle$ is the vacuum state. This state has a potential energy offset U with respect to states where the atoms are separated (Fig. 1a). The pair is unable to decay by converting the potential energy into kinetic energy, as the Bloch band allows a maximum kinetic energy for two atoms given by $8J$, twice its width. The pair can move around the lattice, with both atoms tunnelling to a neighbouring site (Fig. 1b), but the atoms cannot move independently. The stability of repulsively bound pairs is intimately connected with the absence of dissipation, in contrast to solid state lattices, for example, where interactions with phonons typically lead to rapid relaxation.

We obtain experimental evidence for repulsively bound pairs with a sample of ultracold ⁸⁷Rb atoms in a cubic 3D optical lattice with lattice period $a = 415.22$ nm. The key tool used to prepare and observe the pairs is their adiabatic conversion into chemically

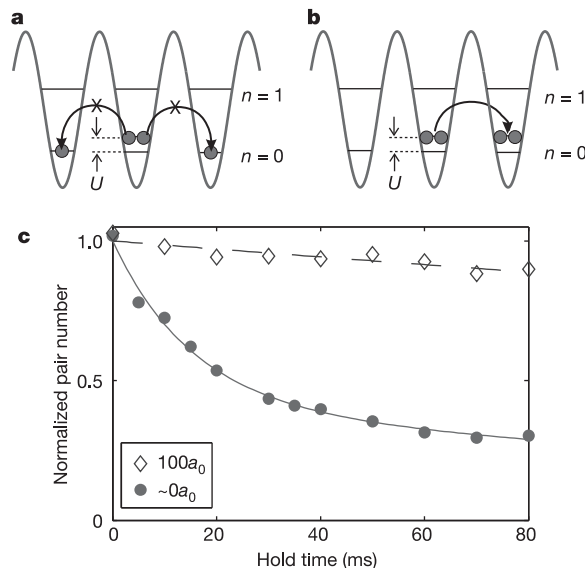


Figure 1 | Atom pairs in an optical lattice. **a**, Repulsive interaction (scattering length $a > 0$) between two atoms sharing a lattice site in the lowest band ($n = 0$) gives rise to an interaction energy U . Breaking up of the pair is suppressed owing to the lattice band structure and energy conservation. **b**, The pair is a composite object that can tunnel through the lattice. **c**, Long lifetime of repulsively bound atom pairs that are held in a 3D optical lattice. The potential depth is $(10 \pm 0.5)E_r$ in one direction and $(35 \pm 1.5)E_r$ in the perpendicular directions. Shown is the remaining fraction of pairs for a scattering length of $100a_0$ (open diamonds; a_0 is the Bohr radius) and a scattering length of about $(0 \pm 10)a_0$ (filled circles) as a function of the hold time. The lines are fitted curves of an exponential (dashed line) and the sum of two exponentials (solid line).

¹Institute for Experimental Physics, ²Institute for Theoretical Physics, University of Innsbruck, A-6020 Innsbruck, Austria. ³Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences, A-6020 Innsbruck, Austria.

bound dimers using a magnetic-field sweep across a Feshbach resonance^{13–20} at 1,007.40 G. The initial state is prepared from a pure sample of Rb₂ Feshbach molecules in the vibrational ground state of the lattice where each lattice site is occupied by not more than a single molecule (see Methods). Sweeping across the Feshbach resonance, we adiabatically dissociate the dimers and obtain a lattice correspondingly filled with 2×10^4 atom pairs, at an effective filling factor of typically 0.3. Away from the Feshbach resonance, the effective interaction between the atoms is repulsive with scattering length $a_s = +100a_0$ (where a_0 is the Bohr radius).

In order to demonstrate the stability of repulsively bound pairs, we lower the lattice potential in one direction from its initial depth of $V_0 = 35E_r$ (corresponding to $J/\hbar \approx 2\pi \times 0.7$ Hz and $U/J \approx 3,700$, where $E_r = 2\pi^2\hbar^2/m\lambda^2$ and m is the mass of the atoms) in 1 ms to a depth of $V_0 = 10E_r$. This increases dramatically the tunnelling rates along this direction to $J/\hbar \approx 2\pi \times 63$ Hz ($U/J \approx 30$), potentially allowing the atom pairs to quickly separate. After a variable hold time we determine the number of remaining pairs. This is done by adiabatically raising the lattice to its full initial depth of $V_0 = 35E_r$, and converting doubly occupied sites to Feshbach molecules with near unit efficiency¹³. A purification pulse¹³ then removes all remaining atoms due to dissociated pairs. Afterwards the molecules are again converted back into atoms, and can then be detected by conventional absorption imaging.

The results of these lifetime measurements are shown in Fig. 1c. For repulsive interaction ($a_s = 100a_0$), the atom pair sample exhibits the remarkably long lifetime of 700 ms (exponential fit). This lifetime is mainly limited by inelastic scattering of lattice photons¹³, and greatly exceeds the calculated time for an atom to tunnel from one site to the next, $2\pi\hbar/(4J) \approx 4$ ms. In contrast, if we turn off the on-site interaction by tuning the scattering length near zero, we observe a much faster decay in the number of doubly occupied sites owing to the rapid diffusion of unbound atoms through the lattice (Fig. 1c). This observation clearly demonstrates that the stability of the pairs is induced by the on-site interaction U .

We can more deeply understand these repulsively bound pairs through an exact solution of the two-particle Lippmann–Schwinger scattering equation based on the Bose–Hubbard model. We write the two-atom wavefunction as $\Psi(\mathbf{x}, \mathbf{y})$, where the positions of the two particles are denoted $\mathbf{x} = \sum_i x_i \mathbf{e}_i$ and $\mathbf{y} = \sum_i y_i \mathbf{e}_i$, with $\mathbf{e}_i$ being the primitive lattice vectors, and x_i, y_i integer numbers. Introducing centre of mass, $\mathbf{R} = (\mathbf{x} + \mathbf{y})/2$, and relative coordinates, $\mathbf{r} = \mathbf{x} - \mathbf{y}$, we can solve the Schrödinger equation with the ansatz $\Psi(\mathbf{x}, \mathbf{y}) = \exp(i\mathbf{K}\mathbf{R})\psi_K(\mathbf{r})$, where $\mathbf{K}$ is the quasi-momentum of the centre of mass motion and $\psi_K(\mathbf{r})$ is the pair wavefunction. We derive two types of solutions (for details see Methods), each of which are eigenstates of $\mathbf{K}$. These states, as illustrated in Fig. 2a, correspond to (1) unbound scattering solutions (shaded area in Fig. 2a), where the two particles move on the lattice, and scatter from each other according to the interaction U , and (2) repulsively bound pairs for

which $\psi_K(\mathbf{r})$ is square integrable. In one and two dimensions, states of repulsively bound pairs always exist for non-zero U , while in three dimensions they exist above a critical value $U_{\text{crit}} \approx 0.5J$.

In this Letter, we focus primarily on the 1D situation, which in the experiment corresponds to a low depth of the lattice along one direction, whilst the lattice in the perpendicular directions remains very deep ($35E_r$). Here the energy of the bound pairs is $E(K) = 2J \left[\sqrt{4(\cos \frac{Ka}{2})^2 + (U/2J)^2} + 2 \right]$. This is plotted in Fig. 2a as the Bloch band of a stable composite object above the continuum of two-particle scattering states. In the limit of strong interaction, $U \gg J$, this reduces to $E(K) \approx 4J + U + (4J^2/U)(1 + \cos Ka)$, which shows that the bound pairs indeed have binding energy of $\sim U$ and hop through the lattice with an effective tunnelling rate $J^2/(\hbar U)$.

Figure 2b shows the pair wavefunctions $\psi_K(r)$ for repulsively bound pairs ($a_s = 100a_0$) in one dimension with $K = 0$, for $U/J = 30$ ($V_0 \approx 10E_r$) and $U/J = 3$ ($V_0 \approx 3E_r$). For large U/J , bound pairs essentially consist of two atoms occupying the same site, whereas for small U/J , the pair is delocalized over several lattice sites. The corresponding quasi-momentum distribution can be found from the Fourier transform $\tilde{\psi}_0(k)$ of the pair wavefunction (Fig. 2c), where k is the relative quasi-momentum. Because $K = 0$, $|\tilde{\psi}_0(k)|^2$ is also equal to the single-particle quasi-momentum distribution. When the two particles are localized on the same site, the quasi-momentum distribution is essentially flat. However, for lower U/J the wavefunction is characteristically peaked at the edges of the Brillouin zone. This occurs because the energy of the repulsively bound state is above that of the continuum, and thus the contribution to the corresponding wavepacket of single-particle quasi-momentum states with higher energy is favoured. In contrast, if we had $U < 0$, the pair would be attractively bound, and would have energy lower than that in the continuum. Thus contributions from the low-energy quasi-momentum states would be favoured, leading to a single peak in the centre of the Brillouin zone. In both cases, the amplitude of the peaks grows with increasing width $4J$ of the Bloch band. In general, the stable bound pairs will not be prepared in a fixed quasi-momentum state K in an experiment, but rather in a superposition of different momentum states. For non-zero K , the peaks in the single-particle quasi-momentum distribution are translated by K , but their strength is also reduced. As a consequence, for typical symmetric distributions of K , the peak at the edge of the Brillouin zone remains present, but is less strong than in the optimal case of vanishing K . We have verified this using many-body numerical simulations, which were performed using time-dependent density-matrix renormalization group methods^{21–23}.

We have experimentally investigated the quasi-momentum distribution of the pairs in various regimes by mapping it onto a spatial distribution, which we measured using standard absorption imaging. For this, we first adiabatically lower the lattice depth in the X

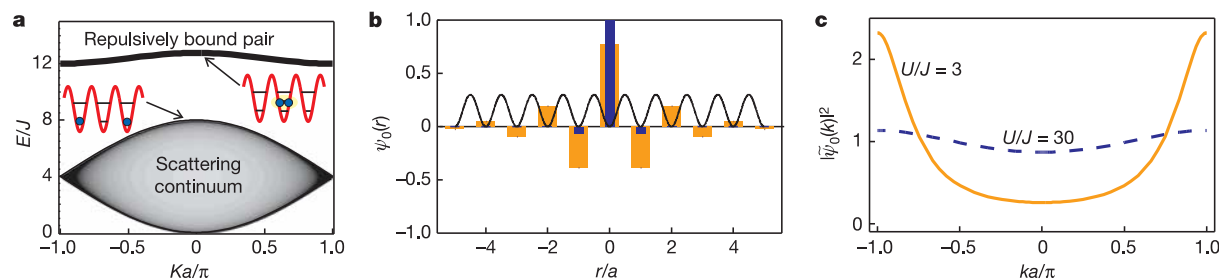


Figure 2 | Atom pair states in one dimension. **a**, Spectrum of energy E of the 1D hamiltonian for $U/J = 8$ ($V_0 \approx 6E_r$) as a function of centre of mass quasi-momentum K . The Bloch band for repulsively bound pairs is located above the continuum of unbound states. The grey level for the shading of the continuum is proportional to the density of states. **b**, The pair wavefunction

$\psi_0(r)$, showing the amplitude at each site with $U/J = 30$ ($V_0 \approx 10E_r$, blue bars) and $U/J = 3$ ($V_0 \approx 3E_r$, orange bars). **c**, The square modulus of the corresponding momentum space wavefunctions $|\tilde{\psi}_0(k)|^2$, which are equivalent to the single-particle momentum distributions, as $K = 0$. Note the characteristic peaks at the edge of the Brillouin zone.

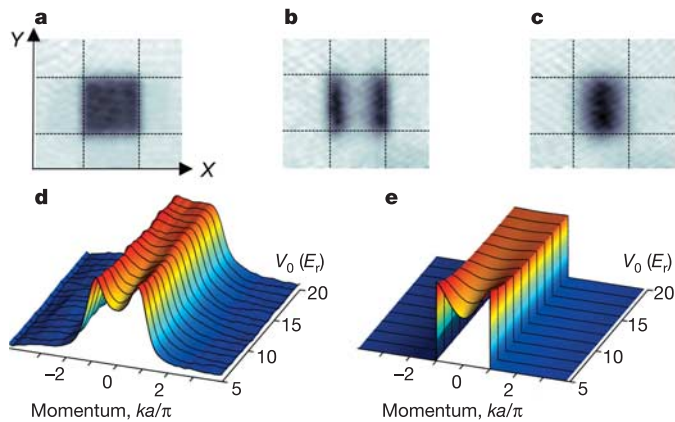


Figure 3 | Quasi-momentum distribution of atoms in the lattice. **a–c**, Absorption images of the atomic distribution after release from the 3D lattice and a subsequent 15-ms time of flight. The horizontal and vertical black lines enclose the first Brillouin zone. **a**, Distribution when lattice sites are occupied by single atoms; **b**, distribution for repulsively bound atom pairs (see text for details); **c**, same as **b** but pairs are attractively bound. **d, e**, The quasi-momentum distribution for pairs in the X direction as a function of lattice depth V_0 , after integration over the Y direction. **d**, Experiment; **e**, numerical calculation. See Methods for a definition of E_r .

direction (Fig. 3a) at a rate of $1.3E_r$ (ms^{-1}) to a pre-chosen height while the lattice depth in the other two directions is kept high ($35E_r$). This will prepare repulsively bound pairs at the chosen lattice depth. We then turn off the lattice rapidly enough for the pair wavefunction not to change, but slowly with respect to the bandgap, so that single-particle quasi-momenta are mapped to real momenta^{24,25}. We have typically employed linear ramps with rates of $0.2E_r \mu\text{s}^{-1}$. The resulting momentum distribution is converted to a spatial distribution after ~ 15 ms time of flight.

Figure 3a–c shows typical measured quasi-momentum distributions that were obtained after adiabatically lowering the lattice depth in the X direction to the lowest values, below $3E_r$. If only empty sites and sites with single atoms are present in the lattice, then the first Brillouin zone is homogeneously filled²⁴ (Fig. 3a). For repulsively bound pairs, the momentum distribution is, in general, peaked at the edges of the first Brillouin zone (Fig. 3b), whereas for attractively bound pairs, it is peaked in the centre of the first Brillouin zone (Fig. 3c). In order to change the interaction between the atoms from repulsive to attractive, we change the scattering length, making use of the Feshbach resonance²⁶ at 1,007.40 G. Figure 3d and e shows the dependence on lattice depth V_0 of the single-particle quasi-momentum distribution for repulsively bound pairs from experiment and numerical simulation, respectively. As expected, the peak structure is more pronounced for lower values of V_0 , and diminishes for larger V_0 . This characteristic is a clear signature of the pair wavefunction for repulsively bound pairs.

We also performed spectroscopic measurements, determining the binding energy from pair dissociation produced by modulating the depth of the lattice at a chosen frequency. On resonance, the modulation allows pairs to release their binding energy. Figure 4a shows the number of remaining pairs as a function of the modulation frequency. This was repeated for a variety of lattice depths V_0 in one direction while keeping the lattice in the other two directions at $35E_r$. The behaviour of the binding energy as a function of the lattice depth provides an additional key signature of repulsively bound pairs. As shown in Fig. 4b, the resonance positions are in good agreement with numerical simulations and essentially coincide with the interaction energy, U .

It is important to note that for sufficiently large U/J , repulsively bound pairs are stable under collisions with each other. This is particularly evident in the limit $U \gg J$ where, by energy arguments,

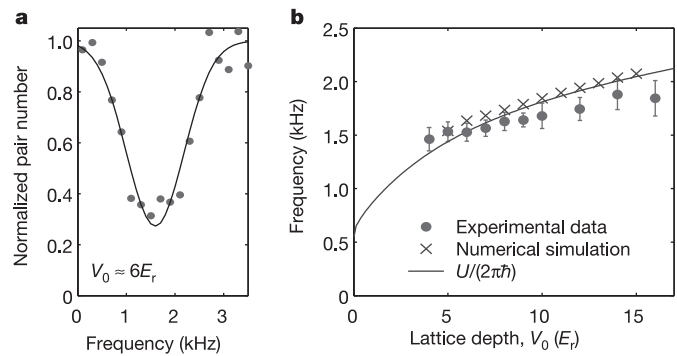


Figure 4 | Modulation spectroscopy of repulsively bound pairs. **a**, Typical resonance dip showing the remaining number of atom pairs as a function of the modulation frequency, for $V_0 \approx 6E_r$. The solid line is a gaussian fit, a choice that was justified by numerical calculations. **b**, Plot showing the measured resonance frequencies (filled circles) as a function of the lattice depth. They show good agreement with numerical simulations (crosses) and also coincide with the on-site collisional energy shift U (line). Experimental error bars correspond to the 95% confidence interval for the gaussian fit parameters of the resonance dips.

the elastic scattering between pairs is the only open channel. This means that even a relatively dense quantum lattice gas of these objects can be long-lived. When the lattice height is lowered so that U/J becomes sufficiently small, it is possible for a certain fraction of the pairs to dissociate by collision with other pairs. In our experiments, we observe the onset of this behaviour for lattice depths lower than $6E_r$, that is, $U/J \approx 9$. The dynamics of the collisions and details of the decay depend crucially on lattice depth and the local density of pairs across the lattice. Further details of these processes will be discussed elsewhere.

In conclusion, we have demonstrated the formation of a novel composite object in an optical lattice: a stable bound state that arises from the lattice band structure and repulsion between the constituents. Although no direct analogue to repulsively bound atomic pairs is known to exist, the formation of a metastable state is reminiscent of trapping light in photonic bandgap materials²⁷, or extended lifetimes of excited atoms in cavity quantum electrodynamics²⁸. In both cases, decay is suppressed by restriction of the accessible light field modes. Stable repulsively bound objects should be viewed as a general phenomenon, and their existence will be ubiquitous in cold atom lattice physics. They also give rise to new potential composites with fermions²⁹ or Bose–Fermi mixtures³⁰, and can be formed in an analogous manner with more than two particles. The stability of repulsively bound objects could thus be the basis of a wealth of new quantum many-body states or phases. In particular, the next experimental step in investigating repulsively bound atomic pairs is the possible realization of a condensate of pairs, together with the means to characterise long-range order in this system.

METHODS

Preparation of pure molecular sample. We use a set-up which was described in detail in ref. 13, starting with a Bose–Einstein condensate of 6×10^5 ^{87}Rb atoms in an Ioffe-type magnetic trap with trap frequencies $\omega_{x,y,z} = 2\pi \times (7, 19, 20)$ Hz. Within 100 ms the Bose–Einstein condensate is adiabatically loaded into the cubic 3D optical lattice which is $35E_r$ deep. After turning off the magnetic trap, we flip the spins of our atoms from their initial state $|F=1, m_F=-1\rangle$ to $|F=1, m_F=+1\rangle$ by suddenly reversing the bias magnetic field of a few gauss. This spin state features a 210-mG-wide Feshbach resonance at 1,007.40 G (ref. 26). By adiabatically ramping over this resonance we convert pairs of atoms in multiply occupied lattice sites into Rb_2 Feshbach molecules. Fast inelastic collisions of molecules within lattice sites and a subsequent combined radio-frequency and optical purification pulse remove all chemically unbound atoms, thus creating a pure molecular sample of about 2×10^4 molecules.

Exact solution for single pair bound state. Within the Bose–Hubbard model (equation (1)), the Schrödinger equation describing two particles in a

homogenous optical lattice takes the form

$$\left[-J\left(\tilde{\Delta}_{\mathbf{x}}^0 + \tilde{\Delta}_{\mathbf{y}}^0\right) + U\delta_{\mathbf{x},\mathbf{y}}\right]\Psi(\mathbf{x},\mathbf{y}) = E\Psi(\mathbf{x},\mathbf{y}) \quad (2)$$

where the vectors $\mathbf{x}$ and $\mathbf{y}$ describe the positions of the two particles as defined in the main text. The operator $\tilde{\Delta}_{\mathbf{x}}^K\Psi(\mathbf{x}) = \sum_{i=1}^d \cos(\mathbf{K}\mathbf{e}_i/2)[\Psi(\mathbf{x} + \mathbf{e}_i) + \Psi(\mathbf{x} - \mathbf{e}_i)] - 2\Psi(\mathbf{x})$ denotes the discrete lattice laplacian with d the dimensionality in the cubic lattice, and $\delta_{\mathbf{x},\mathbf{y}}$ is a Kronecker delta. Writing the wavefunction in relative and centre of mass coordinates $\Psi(\mathbf{x},\mathbf{y}) = \exp(i\mathbf{K}\mathbf{r})\psi_K(\mathbf{r})$, the Schrödinger equation (2) then reduces to a single-particle problem in the relative coordinate

$$\left[-2J\tilde{\Delta}_{\mathbf{r}}^K + E_K + U\delta_{\mathbf{r},0}\right]\psi_K(\mathbf{r}) = E\psi_K(\mathbf{r}) \quad (3)$$

with $E_K = 4J\Sigma_i[1 - \cos(\mathbf{K}\mathbf{e}_i/2)]$ being the kinetic energy of the centre of mass motion.

The short range character of the interaction potential allows for a resummation of the perturbation expansion generated by the corresponding Lippmann–Schwinger equation. We obtain the scattering states

$$\psi^{(+)}(\mathbf{r}) = \exp(i\mathbf{K}\mathbf{r}) - 8\pi J f_E(\mathbf{K}) G_K(E, \mathbf{r}) \quad (4)$$

with scattering amplitude

$$f_E(\mathbf{K}) = -\frac{1}{4\pi} \frac{U/(2J)}{1 - G_K(E, 0)U} \quad (5)$$

where the total energy $E = \epsilon_{\mathbf{K},\mathbf{K}} + E_K$, and $\epsilon_{\mathbf{K},\mathbf{K}} = 4J\Sigma_i \cos(\mathbf{K}\mathbf{e}_i/2)[1 - \cos(\mathbf{K}\mathbf{e}_i)]$. Furthermore, $G_K(E, \mathbf{r})$ denotes the Green's function of the non-interacting problem, which in Fourier space takes the form $\tilde{G}_K(E, \mathbf{k}) = 1/(E - \epsilon_{\mathbf{K},\mathbf{K}} + i\eta)$.

The energy spectrum for these states in one dimension is shown as a function of K by the shaded region in Fig. 2a. In addition, the pole in the scattering amplitude indicates the presence of an additional bound state. The energy E_{bs} of the bound state is determined by $G_K(E_{bs}, 0)U = 1$ and the bound state wavefunction takes the form $\psi^{bs}(\mathbf{r}) = cG_K(E_{bs}, \mathbf{r})$, with c being a normalization factor.

Received 1 April; accepted 18 May 2006.

- Jaksch, D. & Zoller, P. The cold atom Hubbard toolbox. *Ann. Phys.* **315**, 52–79 (2005).
- Bloch, I. Ultracold quantum gases in optical lattices. *Nature Phys.* **1**, 23–30 (2005).
- Fisher, M. P. A., Weichman, P. B., Grinstein, G. & Fisher, D. S. Boson localization and the superfluid insulator transition. *Phys. Rev. B* **40**, 546–570 (1989).
- Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415**, 39–44 (2002).
- Paredes, B. *et al.* Tonks–Girardeau gas of ultracold atoms in an optical lattice. *Nature* **429**, 277–281 (2004).
- Stöferle, T., Moritz, H., Schori, C., Köhl, M. & Esslinger, T. Transition from a strongly interacting 1D superfluid to a Mott insulator. *Phys. Rev. Lett.* **92**, 130403 (2004).
- Kinoshita, T., Wenger, T. & Weiss, D. S. Observation of a one-dimensional Tonks–Girardeau gas. *Science* **305**, 1125–1128 (2004).
- Laburthe Tolra, B. *et al.* Observation of reduced three-body recombination in a correlated 1D degenerate Bose gas. *Phys. Rev. Lett.* **92**, 190401 (2004).
- Fallani, L., Lye, J. E., Guarrera, V., Fort, C. & Inguscio, M. Onset of a Bose-glass of ultra-cold atoms in a disordered crystal of light. Preprint at (<http://arxiv.org/abs/cond-mat/0603655>) (2006).
- Fedichev, P. O., Bijlsma, M. J. & Zoller, P. Extended molecules and geometric scattering resonances in optical lattices. *Phys. Rev. Lett.* **92**, 080401 (2004).
- Ryu, C. *et al.* Raman-induced oscillation between an atomic and a molecular quantum gas. Preprint at (<http://arxiv.org/abs/cond-mat/0508201>) (2005).
- Stöferle, T., Moritz, H., Günter, K., Köhl, M. & Esslinger, T. Molecules of fermionic atoms in an optical lattice. *Phys. Rev. Lett.* **96**, 030401 (2006).
- Thalhammer, G. *et al.* Long-lived Feshbach molecules in a 3D optical lattice. *Phys. Rev. Lett.* **96**, 050402 (2006).
- Donley, E. A., Claussen, N. R., Thompson, S. T. & Wieman, C. E. Atom–molecule coherence in a Bose–Einstein condensate. *Nature* **417**, 529–533 (2002).
- Regal, C. A., Ticknor, C., Bohn, J. L. & Jin, D. S. Creation of ultracold molecules from a Fermi gas of atoms. *Nature* **424**, 47–50 (2003).
- Herbig, J. *et al.* Preparation of a pure molecular quantum gas. *Science* **301**, 1510–1513 (2003).
- Xu, K. *et al.* Formation of quantum-degenerate sodium molecules. *Phys. Rev. Lett.* **91**, 210402 (2003).
- Cubizolles, J., Bourdel, T., Kokkelmans, S. J. J. M. F., Shlyapnikov, G. V. & Salomon, C. Production of long-lived ultracold Li₂ molecules from a Fermi gas. *Phys. Rev. Lett.* **91**, 240401 (2003).
- Dürr, S., Volz, T., Marte, A. & Rempe, G. Observation of molecules produced from a Bose–Einstein condensate. *Phys. Rev. Lett.* **92**, 020406 (2004).
- Koehler, T., Goral, K. & Julienne, P. S. Production of cold molecules via magnetically tunable Feshbach resonances. Preprint at (<http://arxiv.org/abs/cond-mat/0601420>) (2006).
- Vidal, G. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.* **91**, 147902 (2003).
- Daley, A. J., Kollath, C., Schollwöck, U. & Vidal, G. Time-dependent density-matrix renormalization-group using adaptive effective Hilbert spaces. *J. Stat. Mech. Theor. Exp.* P04005 (2004).
- White, S. R. & Feiguin, A. E. Real-time evolution using the density matrix renormalization group. *Phys. Rev. Lett.* **93**, 076401 (2004).
- Greiner, M., Bloch, I., Mandel, O., Hänsch, T. W. & Esslinger, T. Exploring phase coherence in a 2D lattice of Bose–Einstein condensates. *Phys. Rev. Lett.* **87**, 160405 (2001).
- Hecker Denschlag, J. *et al.* A Bose–Einstein condensate in an optical lattice. *J. Phys. B* **35**, 3095–3110 (2002).
- Volz, T., Dürr, S., Ernst, S., Marte, A. & Rempe, G. Characterization of elastic scattering near a Feshbach resonance in ⁸⁷Rb. *Phys. Rev. A* **68**, 010702 (2003).
- Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, Princeton, 1995).
- Berman, P. (ed.) *Cavity Quantum Electrodynamics* (Academic, New York, 1994).
- Hofstetter, W., Cirac, J. I., Zoller, P., Demler, E. & Lukin, M. D. High-temperature superfluidity of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **89**, 220407 (2002).
- Lewenstein, M., Santos, L., Baranov, M. A. & Fehrmann, H. Atomic Bose–Fermi mixtures in an optical lattice. *Phys. Rev. Lett.* **92**, 050401 (2004).

Acknowledgements We thank H. Ritsch for discussions, and M. Theis and S. Schmid for help in setting up the experiment. We acknowledge support from the Austrian Science Fund (FWF) within the Spezialforschungsbereich 15, from the European Union within the OLAQUI and SCALA networks, from the TMR network ‘Cold Molecules’, and from the Tiroler Zukunftsstiftung.

Author Contributions This work is a collaboration between teams of experimental (K.W., G.T., F.L., R.G. and J.H.D.) and theoretical (A.J.D., A.K., H.P.B. and P.Z.) physicists.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.H.D. (Johannes.Denschlag@uibk.ac.at).

Amorphous silica-like carbon dioxide

Mario Santoro^{1,2}, Federico A. Gorelli^{1,2}, Roberto Bini^{1,3}, Giancarlo Ruocco^{2,4}, Sandro Scandolo⁵ & Wilson A. Crichton⁶

Among the group IV elements, only carbon forms stable double bonds with oxygen at ambient conditions. At variance with silica and germania, the non-molecular single-bonded crystalline form of carbon dioxide, phase V, only exists at high pressure^{1–9}. The amorphous forms of silica (a-SiO₂) and germania (a-GeO₂) are well known at ambient conditions; however, the amorphous, non-molecular form of CO₂ has so far been described only as a result of first-principles simulations⁹. Here we report the synthesis of an amorphous, silica-like form of carbon dioxide, a-CO₂, which we call ‘a-carbonia’. The compression of the molecular phase III of CO₂ between 40 and 48 GPa at room temperature initiated the transformation to the non-molecular amorphous phase. Infrared spectra measured at temperatures up to 680 K show the progressive formation of C–O single bonds and the simultaneous disappearance of all molecular signatures. Furthermore, state-of-the-art Raman and synchrotron X-ray diffraction measurements on temperature-quenched samples confirm the amorphous character of the material. Comparison with vibrational and diffraction data for a-SiO₂ and a-GeO₂, as well as with the structure factor calculated for the a-CO₂ sample obtained by first-principles molecular dynamics⁹, shows that a-CO₂ is structurally homologous to the other group IV dioxide glasses. We therefore conclude that the class of archetypal network-forming disordered systems, including a-SiO₂, a-GeO₂ and water, must be extended to include a-CO₂.

High pressure can change substantially the microscopic interactions in condensed phases, and can lead to unexpected transformations. This is even the case in simple molecular systems, where new materials are obtained as a result of dramatic changes of the molecular structure. Among paradigmatic molecular crystals such as those made of CO₂ and N₂, high pressure has been recently shown to lead to a transformation into non-molecular ordered structures^{1,2,10}, and also, in the case of nitrogen, into amorphous materials^{11–13}. Carbon dioxide is the dominant component of the atmospheres of Earth-like planets, is embedded in ice form in the outer planets and asteroids, plays an important role in volcanic and seismic activity¹⁴ and is used as a supercritical solvent in chemical reactions¹⁵. Its high natural abundance and chemical stability can be traced to the presence of a strong π bond, whose strength is however severely affected by pressure¹⁶. The recent discovery of a crystalline ‘super-hard’ phase of carbon dioxide (CO₂-V) composed of CO₄ tetrahedra indicates that the strong molecular bond is replaced, at high pressure, by an extended network of C–O single bonds, structurally similar to many ordered SiO₂ polymorphs. Naturally, the question therefore arises if there is the potential for non-molecular glassy phases and/or stishovite-like structures in carbon dioxide. The existence of an amorphous extended form of CO₂ would extend the class of network-forming systems, to which a-SiO₂

and water belong, to include a material with very high bond stiffness and thus probably remarkable mechanical properties. Stishovite-like ordered structures, with octahedral carbon coordination by oxygen, have been already considered in first-principles calculations^{8,9}. They were predicted to be very hard and ionic, and to be stable over a number of tetrahedral polymorphs above 400 GPa (ref. 8).

We studied modifications of carbon dioxide by conducting five separate experiments close to the recently determined⁴ transformation boundary from CO₂-III to CO₂-V, that is, at 40–76 GPa and 300–680 K (see Fig. 1b inset), by using an externally heated diamond anvil cell. Phase III was compressed at room temperature to between 40 and 48 GPa until a new, broad, infrared band (B_{IR} in Fig. 1a), composed of at least two components and centred at about 1,250 cm^{–1}, emerged. This strong band, which saturates the absorption when the transformation is enhanced by increasing temperature, is not due to molecular CO₂ and therefore indicates a chemical transformation of the sample. Another band not belonging to molecular CO₂, A_{IR}, centred around 730 cm^{–1}, is also observed in the transformed sample. The width of band B_{IR} is a preliminary hint of disorder in the structure. The almost complete disappearance of the infrared molecular lines, Fig. 1, confirms that the transformation into a non-molecular phase (a-CO₂) is complete at 564 K and 64 GPa.

The temperature quenched a-CO₂ samples are transparent (see inset of Fig. 2), and were investigated by means of infrared (IR) and Raman spectroscopy, and X-ray diffraction techniques. At variance with a study of CO₂ at very high pressures and temperatures (above 2,000 K)³, where the presence of molecular oxygen, carbon clusters, and large spatial inhomogeneities were reported in the recovered samples, we did not find evidence of decomposition, probably owing to the milder temperature conditions of our experiments. We also did not find traces of carbonyl groups. The comparison of the IR spectrum of a-CO₂ with those of different silica polymorphs (Fig. 2) such as a-SiO₂, quartz, coesite and stishovite suggests a silica-like character for our non-molecular extended phase. When compressed to 25–35 GPa, both quartz and coesite amorphize¹⁷ and their IR spectra, along with that of a-SiO₂ compressed to similar pressures, tend to resemble that of stishovite^{18,19}. By shifting the frequencies according to the reduced mass ratio between Si–O and C–O pairs, the overall spectral shapes of all dense silica polymorphs are found to resemble closely that of a-CO₂. The shifted silica band at about 650 cm^{–1} appears to be related to band A_{IR} of a-CO₂, which we therefore assign to a deformation of the carbon–oxygen framework. The broad feature lying between 800 and 1,700 cm^{–1} in the SiO₂ shifted spectrum can be related to band B_{IR} of a-CO₂; this feature corresponds, in SiO₂, to stretching and O–Si–O bending vibrations within SiO₆ octahedra (between 800 and 1,300 cm^{–1}), and to SiO₄ tetrahedral stretching vibrations (above 1,300 cm^{–1}). On the basis

¹LENS, European Laboratory for Non-linear Spectroscopy and INFN, Via N. Carrara 1, I-50019 Sesto Fiorentino, Firenze, Italy. ²CRS-SOFT-INFN-CNR, c/o Università di Roma “La Sapienza”, I-00185 Roma, Italy. ³Dipartimento di Chimica dell’Università di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Firenze, Italy. ⁴Dipartimento di Fisica, Università di Roma “La Sapienza”, I-00185 Roma, Italy. ⁵The Abdus Salam International Centre for Theoretical Physics (ICTP) and INFN/Democritos National Simulation Center, 34014 Trieste, Italy. ⁶European Synchrotron Research Facility, BP 220, F38043 Grenoble, France.

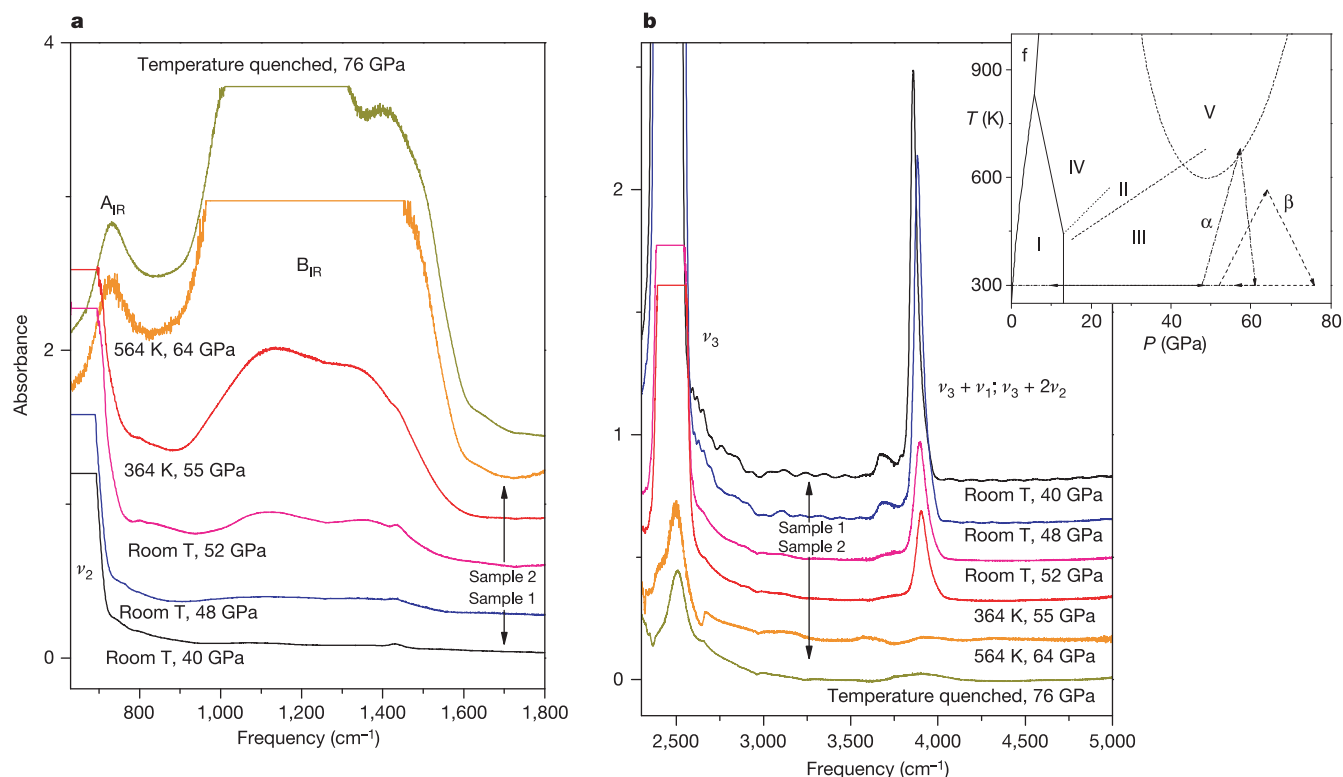
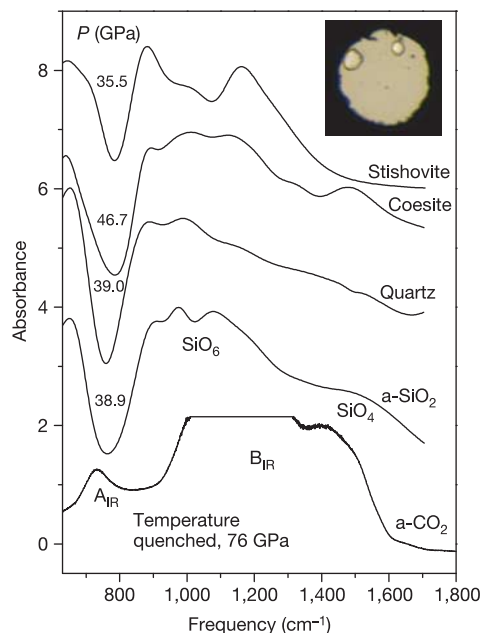


Figure 1 | Infrared spectra showing the transformation of the molecular solid, CO₂-III, into a-CO₂ for two different samples. In **a**, the initial spectrum, at room temperature and 40 GPa, is dominated by the saturating ν_2 band of CO₂ below 800 cm⁻¹. A new broad band, B_{IR}, formed by at least two components, appears on increasing pressure from 40 to 52 GPa. At 564 K, the intensity of band B_{IR} saturates the absorption in the spectral region between 965 and 1,450 cm⁻¹, and another new band, A_{IR}, is observed around 730 cm⁻¹. Pressure changes were not compensated along the heating cycle. In **b**, the spectral region of the saturating molecular ν_3 line and of the $\nu_3 + \nu_1$ and $\nu_3 + 2\nu_2$ combination bands is shown in reverse order versus temperature, for the sake of visual clarity. Just a remnant of the ν_3 line is evident in spectra of the transformed sample. In the inset, we report the

state-of-the-art phase diagram of CO₂ (refs 4, 24): I, II, III, IV, molecular crystalline phases; f, fluid phase; V, non-molecular crystal. Dashed lines are kinetic barriers. The apparent quadruple point that seems to connect phases I, IV, II and III does not imply any thermodynamic paradox, as the line between CO₂-II and -III is a kinetic barrier rather than a phase boundary. Arrows schematically indicate the *P*-*T* paths of our five separate experiments: paths α (dot dashed) and β (dashed). Spectra in this figure are relative to path β , while similar IR spectra were obtained in synthesizing a-CO₂ along path α . Spectra were measured by means of a Fourier transform infrared spectrometer (Bruker IFS-120 HR), equipped with an optical beam condenser system²⁵. Starting sample thickness and diameter at ambient pressure were about 40 and 100 μ m, respectively.

Figure 2 | Comparison of a-CO₂ IR spectrum with mass-shifted spectra. Shown are the IR spectrum of a-CO₂ and mass-shifted spectra of a-SiO₂, pressure amorphized quartz and coesite, and stishovite^{18,19}. Mass shift was obtained by multiplying the frequency scale relative to SiO₂ by $(\mu_{\text{Si-O}}/\mu_{\text{C-O}})^{0.5}$, where $\mu_{\text{Si-O}}$ and $\mu_{\text{C-O}}$ are reduced masses of the Si-O and C-O pair, respectively. Inset, typical temperature quenched a-CO₂ sample (*P* = 61 GPa, sample diameter about 80 μ m; two ruby chips are close to the gasket). The broad feature lying between 800 and 1,700 cm⁻¹ in the SiO₂ shifted spectra corresponds to stretching and O-Si-O bending vibrations within SiO₆ octahedra (between 800 and 1,300 cm⁻¹), and to SiO₄ tetrahedral stretching vibrations (above 1,300 cm⁻¹).



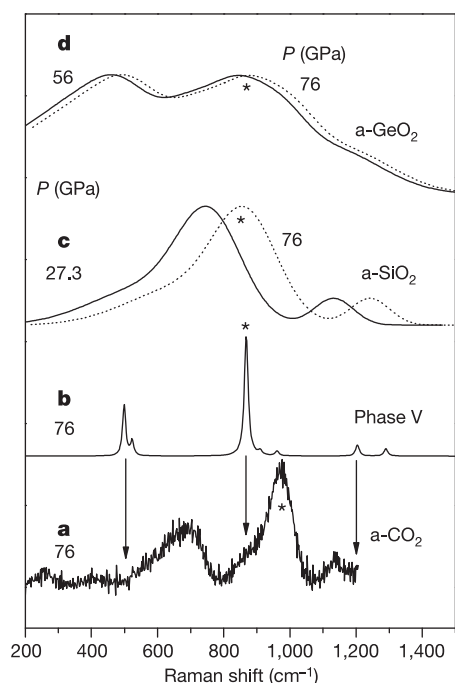


Figure 3 | Comparison of a-CO₂ Raman spectrum with phase V spectra and mass-shifted spectra. **a**, Spectrum of temperature quenched a-CO₂ at 76 GPa, as obtained along the synthesis path β (Fig. 1b inset); **b**, pattern fitted to spectra of temperature quenched phase V (ref. 4), measured up to 66 GPa, and pressure shifted to 76 GPa. Arrows point out the correspondence of phase V lines to band edges of a-CO₂. **c**, **d**, Mass (continuous lines) and mass and pressure (dotted lines) shifted spectra of a-SiO₂ and a-GeO₂, respectively; spectra of glassy SiO₂ and GeO₂ are from refs 20 and 21, respectively. Stars indicate bands assigned to symmetric X–O–X (X = C, Si or Ge) stretching modes. The shifted spectrum of a-GeO₂ exhibits a broad feature centred around 470 cm^{−1}, which was suggested to be related to the high pressure octahedral modification of amorphous germania²¹. Raman measurements on a-CO₂ were performed by using the 752.5 nm line of a Kr⁺ laser as excitation source. The backscattering geometry was used, and the unpolarized signal was detected by a single monochromator (Acton/SpectraPro 2500i) equipped with a CCD detector (Princeton Instruments Spec-10:100BR).

of the analogy between the IR spectra of CO₂ and SiO₂, we infer that a-CO₂ and dense silica have similar local structures.

The Raman spectrum of a-CO₂ is characterized by broad features (Fig. 3), which confirm the amorphous character of the new phase. In particular, we find an asymmetric band at around 690 cm^{−1}, a narrower band at 970 cm^{−1} with a shoulder at 870 cm^{−1} and a weak band at 1,135 cm^{−1}. The spectrum of crystalline phase V (ref. 4) exhibits much narrower lines, shifted slightly with respect to that of a-CO₂. The Raman lines of CO₂-V are close to the a-CO₂ band edges, so we speculate that the broad bands observed in a-CO₂ reflect vibrational density-of-states features of CO₂-V. This also supports the notion that a-CO₂ is the glassy counterpart of phase V. Comparison with the Raman spectra of a-SiO₂ (ref. 20) and a-GeO₂ (ref. 21), available up to 27.3 and 56 GPa, respectively, points to the homology between those glasses and a-CO₂. Spectra of a-Si(Ge)O₂ were properly shifted by considering the different masses (as for IR spectra) and pressures (according to the measured pressure shifts) with respect to those of a-CO₂. The shifted spectra of a-SiO₂ and a-GeO₂ exhibit prominent broad bands peaked at around 845 and 880 cm^{−1}, respectively, corresponding to symmetric Si(Ge)–O–Si(Ge) stretching modes that overlap with the spectral region of the C–O–C stretching modes of phase V and a-CO₂ (peaks at 868 and 970 cm^{−1}, respectively).

As a final test of structural disorder in the new material, we

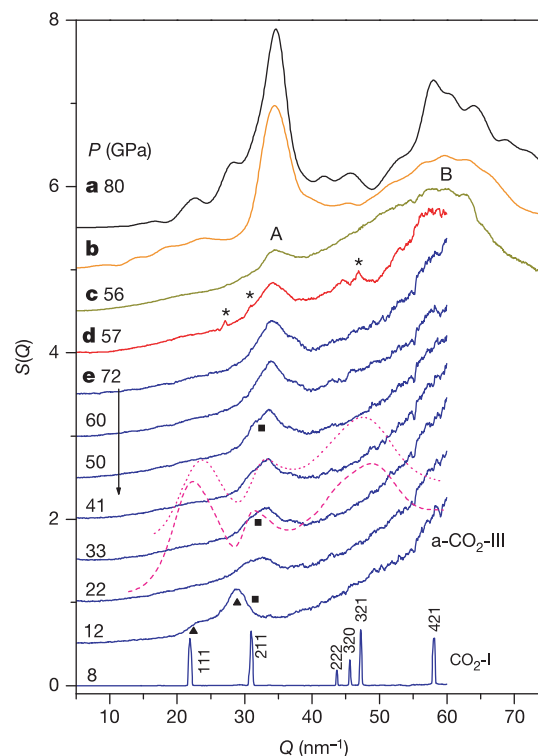


Figure 4 | Static structure factor, $S(Q)$, of a-CO₂. **a**, **b**, Theoretical $S(Q)$ for CO₄ and higher (up to octahedral) carbon coordination, respectively; **c**, **d**, temperature quenched samples synthesized along path α (see Fig. 1b), and **e**, along path β . The structure factor of sample **e** has been measured upon decreasing pressure. At about 16 GPa, a-CO₂ transforms into a new amorphous material, a-CO₂-III, that is suggested to be the glassy counterpart of phase III (see text). Below 10 GPa, the amorphous material recrystallizes into phase I (dry ice), whose peaks are indexed. The shorter Q range relative to sample **d** and **e** was due to the smaller access angle of the cell. The experimental $S(Q)$ of a-SiO₂ is also reported at 42 GPa (dotted line) and 28 GPa (dashed line)²². Symbols indicate diffraction peaks of: (stars) rhenium gasket, (squares) phase V² (the closest line to peak A of a-CO₂), and (triangles) phase III²³. Angle dispersive X-ray diffraction patterns have been measured at the ID27 beam line of the European Synchrotron Research Facility, with monochromatic beam ($\lambda = 0.3738$ Å), and image plate detection. The nominal size of the focal spot was equal to $6 \times 6 \mu\text{m}^2$. The diffraction patterns were analysed and integrated using the FIT2D computer code²⁶. The structure factor $S(Q)$ was obtained as: $S(Q) = [I(Q) - I_b(Q)]/Nf^2$, where $I(Q)$ and $I_b(Q)$ are the polarization-corrected total and empty-cell X-ray diffraction intensities, respectively, N is a normalization factor, and f is the CO₂ molecular form factor. See Supplementary Information for further details. Pattern **a** was obtained by means of *ab initio* molecular dynamics simulations of CO₂, after compression of phase III to 80 GPa at 2,000 K, where a-CO₂ was found to contain only tetrahedral sites. Pattern **b** was obtained by classical molecular dynamic simulations of a-SiO₂, brought to 28 GPa, which we find to contain similar proportions of four-, five-, and six-fold coordinated Si (ref. 27), by scaling its volume to that of simulated a-CO₂, and by artificially replacing the atomic form factor of Si with that of C.

measured its static structure factor, $S(Q)$ (Fig. 4), where Q is the exchanged wave-vector. The $S(Q)$ at 56 GPa is dominated by a broad peak centred around 34.3 nm^{−1} (peak A), which corresponds to a length scale $2\pi/Q \approx 1.83$ Å, and a much broader peak around 58 nm^{−1} (peak B). The theoretical $S(Q)$, calculated on a fully tetrahedral atomic structure obtained by first-principles molecular dynamics⁹, reproduces both experimental peaks A and B, although peak A exhibits a higher intensity in the theoretical pattern. However, similar agreement with the measured $S(Q)$ can be obtained by considering a theoretical sample containing carbon in higher coordination (see Fig. 4 legend). The similarity between the structure

factors calculated with different local environments (from tetrahedral to octahedral) shows that, even though the first-principles molecular dynamics simulations give a fully tetrahedral phase for a-CO₂, the measured structure factor cannot be used to discriminate between differently coordinated local atomic environments.

Network-forming disordered systems like a-SiO₂ are characterized by a peculiar first diffraction peak at considerably lower *Q* than expected on the basis of nearest-neighbour atomic distances, which has been attributed to the existence of correlation over medium length scales. It is thus interesting to compare the *S(Q)* of a-CO₂ with that of a-SiO₂ (ref. 22) to check if a similar peak exists in a-CO₂. At 42 GPa, the length scale related to peak A in a-SiO₂ (1.86 Å) compares well with the expected O–O distance at that pressure, while the peak at lower *Q* corresponds to a length approximately 30% longer. Peak A in a-CO₂ corresponds also to the theoretical O–O distance, but no peak can be found at lower *Q*. However this is not a consequence of lack of medium range order, but rather of a sort of cancellation among the partial structure factors. As a matter of fact, if we replace artificially the atomic form factor of C used to calculate *ab initio* the *S(Q)*, with that of Si, we recover a peak at lower *Q* similarly to a-SiO₂.

We notice that peak A is very close to the strongest peak of crystalline CO₂-V (ref. 2), whose intensity has been suggested to be enhanced by disorder², indicating once again that a-CO₂ is the amorphous counterpart of phase V. The *S(Q)* was also measured upon decreasing pressure, showing a positive pressure shift of peak A ($\Delta Q = 1.63 \text{ nm}^{-1}$ for $\Delta P = 50 \text{ GPa}$), consistent with pressure induced densification. The pressure shift is, however, very small and parallels that of the strongest CO₂-V peak², suggesting that the two materials have similar and very small compressibilities. At about 16 GPa a-CO₂ transforms into a new amorphous material, whose diffraction features appear to be related to the strongest Bragg peaks of phase III (ref. 23). We therefore suggest this to be the amorphous molecular counterpart of phase III. Below 10 GPa, phase I is formed by recrystallization of the amorphous material.

With the discovery of 'a-carbonia', a new, extended analogy emerges between the structures of CO₂ and those of the known SiO₂/GeO₂ polymorphs. CO₂ transforms under pressure into a three-dimensional amorphous network, similar to, but most probably harder than, a-SiO₂ and a-GeO₂. Amorphous carbonia is the disordered parent of tetrahedral quartz-like CO₂-V, though some carbon sites having higher coordination, at least in the sense of coordination defects, may also be included in such a material. This point seems to be specifically supported by the general agreement between IR spectra of a-CO₂ and those of densified a-SiO₂/GeO₂ systems, which are known to be composed by tetrahedral up to octahedral sites. The X-ray diffraction analysis is not conclusive in this respect, and further investigation is needed to definitely assess the possible existence of high coordination sites in carbonia. Our findings demonstrate that non-molecular CO₂ is the stable form of the important compound CO₂ at high pressure and room temperature, as already suggested on the basis of thermodynamic arguments⁴. We hope that the discovery of a-carbonia will stimulate further studies on the high-pressure phase diagram of CO₂, which could extend our understanding of network-forming systems and strengthen the notion of pressure-induced amorphous states.

Received 23 September 2005; accepted 9 May 2006.

1. Iota, V., Yoo, C. S. & Cynn, H. Quartzlike carbon dioxide: an optically nonlinear extended solid at high pressures and temperatures. *Science* **283**, 1510–1513 (1999).

2. Yoo, C. S. *et al.* Crystal structure of carbon dioxide at high pressure: "superhard" polymeric carbon dioxide. *Phys. Rev. Lett.* **83**, 5527–5530 (1999).
3. Tschanner, O., Mao, H. K. & Hemley, R. J. New transformations of CO₂ at high pressures and temperatures. *Phys. Rev. Lett.* **87**, 075701 (2001).
4. Santoro, M., Lin, J. F., Mao, H. K. & Hemley, R. J. In situ high P-T Raman spectroscopy and laser heating of carbon dioxide. *J. Chem. Phys.* **121**, 2780–2787 (2004).
5. Dong, J., Tomfohr, J. K. & Sankey, O. F. Non-molecular carbon dioxide (CO₂) solids. *Science* **287**, 11a (2000).
6. Dong, J., Tomfohr, J. K. & Sankey, O. F. Rigid intertetrahedron angular interaction of nonmolecular carbon dioxide solids. *Phys. Rev. B* **61**, 5967–5971 (2000).
7. Dong, J. *et al.* Investigation of hardness in tetrahedrally bonded nonmolecular CO₂ solids by density-functional theory. *Phys. Rev. B* **62**, 14685–14689 (2000).
8. Holm, B., Ahuja, R., Belonoshko, A. & Johansson, B. Theoretical investigation of high pressure phases of carbon dioxide. *Phys. Rev. Lett.* **85**, 1258–1261 (2000).
9. Serra, S., Cavazzoni, C., Chiarotti, G. L., Scandolo, S. & Tosatti, E. Pressure-induced solid carbonates from molecular CO₂ by computer simulation. *Science* **284**, 788–790 (1999).
10. Eremets, M. I., Gavriluk, A. G., Trojan, I. A., Dzivenko, D. A. & Boehler, R. Single-bonded cubic form of nitrogen. *Nature Mater.* **3**, 558–563 (2004).
11. Goncharov, A. F., Gregoryanz, E., Mao, H. K., Liu, Z. & Hemley, R. J. Optical evidence for a nonmolecular phase of nitrogen above 150 GPa. *Phys. Rev. Lett.* **85**, 1262–1265 (2000).
12. Gregoryanz, E., Goncharov, A. F., Hemley, R. J. & Mao, H. K. High pressure amorphous nitrogen. *Phys. Rev. B* **64**, 052103 (2001).
13. Eremets, M. I., Hemley, R. J., Mao, H. K. & Gregoryanz, E. Semiconducting nonmolecular nitrogen up to 240 GPa and its low pressure stability. *Nature* **411**, 170–174 (2001).
14. Miller, S. A. *et al.* Aftershocks driven by high-pressure CO₂ source at depth. *Nature* **427**, 724–727 (2004).
15. Kendall, J. L., Canelas, D. A., Young, J. L. & DeSimone, J. M. Polymerizations in supercritical carbon dioxide. *Chem. Rev.* **99**, 543–563 (1999).
16. Schettino, V., Bini, R., Ceppatelli, M., Ciabini, L. & Citroni, M. Chemical reactions at very high pressure. *Adv. Chem. Phys.* **11**, 105–242 (2005).
17. Hemley, R. J., Jephcoat, A. P., Mao, H. K., Ming, L. C. & Manghnani, M. H. Pressure-induced amorphization of crystalline silica. *Nature* **334**, 52–54 (1988).
18. Williams, Q. & Jeanloz, R. Spectroscopic evidence for pressure-induced coordination changes in silicate glasses and melts. *Science* **239**, 902–905 (1988).
19. Williams, Q., Hemley, R. J., Kruger, M. B. & Jeanloz, R. High-pressure infrared spectra of α -quartz, coesite, stishovite and silica glass. *J. Geophys. Res.* **98**, 22157–22170 (1993).
20. Hemley, R. J., Mao, H. K., Bell, P. M. & Mysen, B. O. Raman spectroscopy of SiO₂ glass at high pressure. *Phys. Rev. Lett.* **57**, 747–750 (1986).
21. Durben, D. J. & Wolf, G. H. Raman spectroscopic study of the pressure-induced coordination change in GeO₂ glass. *Phys. Rev. B* **43**, 2355–2363 (1991).
22. Meade, C., Hemley, R. J. & Mao, H. K. High-pressure x-ray diffraction of SiO₂ glass. *Phys. Rev. Lett.* **69**, 1387–1390 (1992).
23. Aoki, K., Yamawaki, H., Sakashita, M., Gotoh, Y. & Takemura, K. Crystal structure of the high-pressure phase of solid CO₂. *Science* **263**, 356–358 (1994).
24. Gorelli, F. A., Giordano, V. M., Salvi, P. R. & Bini, R. Linear carbon dioxide in the high-pressure high-temperature crystalline phase IV. *Phys. Rev. Lett.* **93**, 205503 (2004).
25. Gorelli, F. A., Ulivi, L., Santoro, M. & Bini, R. The ϵ phase of solid oxygen: evidence of an O₄ molecule lattice. *Phys. Rev. Lett.* **83**, 4093–4096 (1999).
26. Hammersley, A. P., Svensson, S. O., Hanfland, M., Fitch, A. N. & Häusermann, D. Two-dimensional detector software: from real detector to idealised image or two-theta scan. *High Press. Res.* **14**, 235–248 (1996).
27. Liang, Y., Miranda, C. R. & Scandolo, S. Mechanical strength and coordination defects in compressed SiO₂ glass. *Phys. Rev. Lett.* (submitted).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank F. Sette and M. Mezouar for discussions and the ESRF for provision of beamtime at ID27. S.S. thanks Y. Liang for preparing the molecular dynamics configurations. This work was supported by the European Community and the Ente Cassa di Risparmio di Firenze.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (santoro@lens.unifi.it) or F.A.G. (gorelli@lens.unifi.it).

Control of four stereocentres in a triple cascade organocatalytic reaction

Dieter Enders¹, Matthias R. M. Hüttel¹, Christoph Grondal¹ & Gerhard Raabe¹

Efficient and elegant syntheses of complex organic molecules with multiple stereogenic centres continue to be important in both academic and industrial laboratories¹. In particular, catalytic asymmetric multi-component ‘domino’ reactions, used during total syntheses of natural products and synthetic building blocks, are highly desirable^{2,3}. These reactions avoid time-consuming and costly processes, including the purification of intermediates and steps involving the protection and deprotection of functional groups, and they are environmentally friendly and often proceed with excellent stereoselectivities^{4,5}. Therefore, the design of new catalytic and stereoselective cascade reactions is a continuing challenge at the forefront of synthetic chemistry⁶. In addition, catalytic cascade reactions can be described as biomimetic, as they are reminiscent of tandem reactions that may occur during biosyntheses of complex natural products^{7,8}. Here we report the development of an asymmetric organocatalytic triple cascade reaction for the synthesis of *tetra*-substituted cyclohexene carbaldehydes. This three-component domino reaction proceeds by way of a catalysed Michael/Michael/aldol condensation sequence affording the products with good to moderate yields (25–58 per cent). During this sequence, four stereogenic centres are formed with high diastereoselectivity and complete enantioselectivity. In addition, variation of the starting materials can be used to obtain diverse polyfunctional cyclohexene derivatives, which can be used as building blocks in organic synthesis.

During the past few years, the field of asymmetric catalysis, previously dominated by biocatalysis and metal catalysis, has been complemented by organocatalysis using small organic molecules as a

third powerful tool. Organocatalysts are usually non-toxic, highly efficient and selective, readily available, metal-free and robust, explaining the growing interest in their use for organic synthesis^{9–11}.

In particular, secondary amines—capable of both enamine as well as iminium catalysis—can be used to design ‘domino’ processes¹². Recently, List¹³, MacMillan¹⁴ and Jørgensen¹⁵ showed examples of this property by merging first iminium and then enamine activation. In contrast, we embarked on a reverse strategy using enamine activation of the first substrate to start a triple cascade. The reaction sequence that we have devised is depicted in Fig. 1.

This catalytic cascade is a three component reaction comprising a linear aldehyde **A**, a nitroalkene **B**, an α,β -unsaturated aldehyde **C** and a simple chiral secondary amine, which is capable of catalysing each step of this triple cascade. In the first step (Fig. 2), the catalyst (*S*)-**1** activates component **A** by enamine formation, which then selectively adds to the nitroalkene **B** in a Michael-type reaction¹⁶. The following hydrolysis liberates the catalyst, which is now able to form the iminium ion of the α,β -unsaturated aldehyde **C** to accomplish the conjugate addition with the nitroalkane **3**¹⁷. In the subsequent third step, a further enamine activation of the proposed intermediate **4** leads to an intramolecular aldol condensation via **5**. Hydrolysis returns the catalyst for further cycles and releases the desired *tetra*-substituted cyclohexene carbaldehyde **2**.

It is well known that nitroalkenes **B**¹⁸ are among the most reactive Michael acceptors, explaining the chemoselectivity of the first step of the catalytic cycle. Therefore the enamine of **A** reacts much faster with **B** than with the α,β -unsaturated aldehyde **C**. Once the Michael

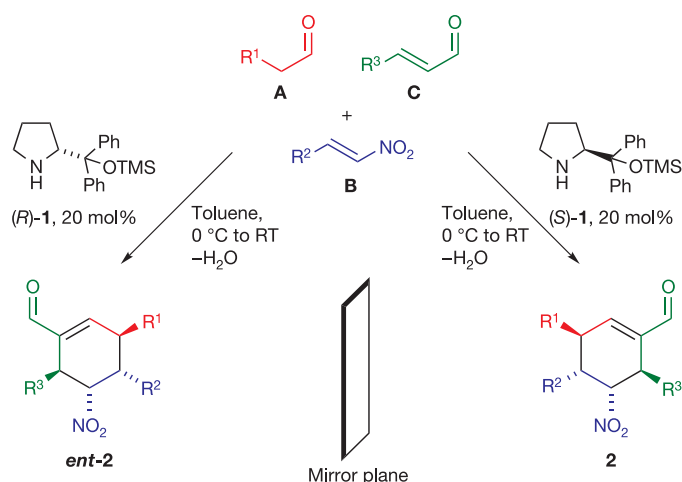


Figure 1 | Asymmetric, organocatalytic three component multistep reaction cascade. RT, room temperature.

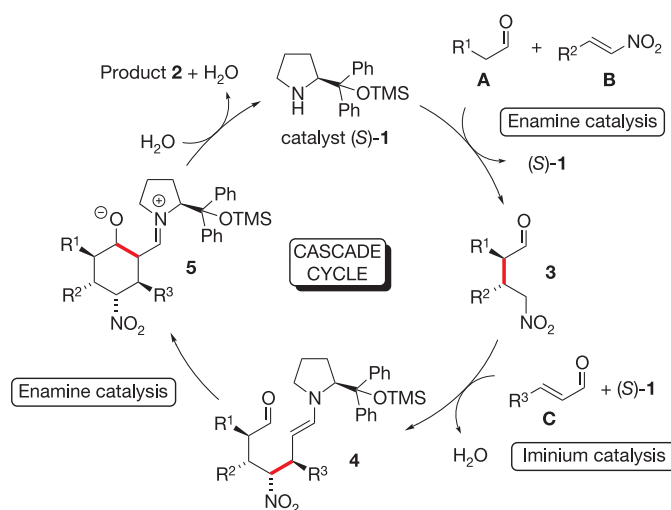


Figure 2 | Proposed catalytic cycle of the triple cascade. The newly formed bond of each step is shown in red for clarity.

¹Institute of Organic Chemistry, RWTH Aachen University, Landoltweg 1, D-52074 Aachen, Germany.

Table 1 | Yield and stereoselectivities of reported reactions

2	R ¹	R ²	R ³	Yield* (%)	d.r.† (%)	d.e., e.e.‡ (%)
a	Me	Ph	Ph	40 (70)	7.8:2.2	>99
b	Me	<i>o</i> -ClPh	Ph	51 (61)	8.4:1.6	>99
c	Me	<i>p</i> -MeOPh	Ph	38 (65)	8.3:1.7	>99
d	Me	Piperonyl	Ph	39 (64)	8.7:1.3	>99
e	Et	Ph	Ph	58 (60)	8.0:2.0	>99
f	<i>i</i> -Pr	Ph	Ph	56 (62)	7.9:2.1	>99
g	Bn	Ph	Ph	38 (45)	8.9:1.1	>99
h	CH ₂ OTBS	Ph	Ph	54 (62)	9.9:0.1	99
i	Me	Ph	H	50 (58)	8.6:1.4	>99
j	Me	Ph	Me	25 (32)	6.8:3.2	>99
k	Me	Ph	<i>n</i> -Bu	29 (33)	8.0:2.0	>99
l	Me	5-Methyl-furan-2-yl	Ph	37 (57)	8.9:1.1	99

Properties of the reported chemo-, diastereo- and enantioselective three component domino reaction affording the tetra-substituted cyclohexene carbaldehydes 2.

* Yield of the isolated main diastereomer (in parenthesis: yield of the crude main diastereomer determined by gas chromatography).

† The diastereomeric ratio (d.r.) was determined by gas chromatography mass spectrometry. The minor diastereomer was determined as the 5-epimer of 2 (Supplementary Information).

‡ The diastereomeric and enantiomeric excess (d.e., e.e.) was determined by HPLC on a chiral stationary phase (Supplementary Information).

§ The reaction was carried out on a 20 mmol scale using only 10 mol% of the catalyst without any loss of the enantiomeric purity and slightly lower yield (37%). 2.4 g of 2a were isolated.

adduct 3 is formed, the following steps are so quick that the intermediates 4 and 5 could not be detected by gas chromatography measurements (see Supplementary Fig. 2). The final product 2, also an α,β -unsaturated aldehyde, is sterically too hindered for a further Michael addition compared to C.

After optimization of the reaction conditions, we carried out the domino reaction in toluene between 0 °C and room temperature, and the reaction was finished after 16–24 hours. Nearly stoichiometric amounts of all three components were used (A:B:C = 1.20:1.00:1.05). The best results concerning yield and selectivity were obtained with 20 mol% of the OTMS-protected (TMS = trimethylsilyl) diphenylprolinol ((S)-1) which is easily available from the natural amino acid (S)-proline^{19,20}. Employing the (R)-proline derived catalyst (R)-1 affords the enantiomeric products *ent*-2. The fact that the residues R¹–R³ of the precursors A, B and C can be varied (see Table 1) demonstrates the high flexibility of our approach. Use of automation techniques may deliver a huge library of different derivatives (for example, using ten different substrates for each compound A, B and C may provide 1,000 derivatives of 2).

As shown in Table 1, R¹ of component A can bear simple to demanding residues as well as valuable functional groups (2h). At present, R² is limited to aromatic and heteroaromatic (2i) substituents. Efforts to use aliphatic nitroalkenes B need further optimization; only traces of the cyclohexene carbaldehydes 2 can be obtained so far. The residue R³ of component C allows the broadest diversity. Aliphatic as well as aromatic moieties are tolerated. Furthermore, acrolein (R³ = H) can be used, affording *tri*-substituted cyclohexene carbaldehydes (2i). The best yields were obtained with aromatic substituents R² and R³ (38–58%). The replacement of R³ by aliphatic residues led to lower yields (25% and 29%), whereas sterically demanding aldehydes A had less influence on the yield. In contrast, the variation of the residues had only a small impact on the diastereoselectivity (diastereomeric ratio d.r. = 6.8:3.2–9.9:0.1), and we were pleased that in all cases complete enantiocontrol (enantiomeric excess e.e. $\geq$ 99%) was obtained.

This cascade reaction generates four stereogenic centres, and theoretically could give rise to 2⁴ = 16 different stereoisomers. We note that in fact this asymmetric cascade forms enantioselectively just two diastereomers, which are easy to separate by flash chromatography. The reason for the high stereoselectivity is the first Michael addition, which is known to proceed with high diastereo- and enantioselectivity^{21,22}; clearly this selectivity is kept or enhanced in the second step via a sterically favourable interaction between the iminium species and nitroalkane 3¹⁷. The relative and absolute

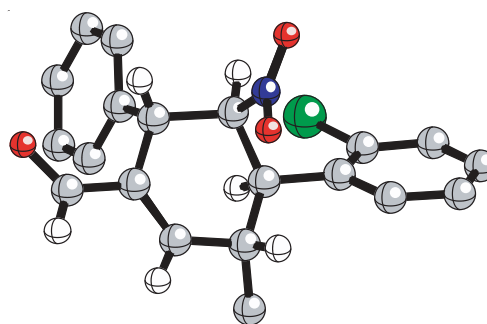


Figure 3 | X-ray structure of 2b. Colour coding: white, hydrogen; grey, carbon; blue, nitrogen; red, oxygen; green, chlorine.

configuration of the cyclohexene carbaldehyde 2 was determined by ¹H-NMR nuclear Overhauser effect (NOE) experiments and X-ray crystallography (Fig. 3). The relative and absolute configurations are in agreement with respective related organocatalytic conjugate additions.

We have developed a chemo-, diastereo- and enantioselective three component domino reaction, accomplished with the readily available proline derived organocatalyst 1 and low cost, simple starting materials, leading to tetra-substituted cyclohexene carbaldehydes 2. The four stereogenic centres are generated in three consecutive carbon–carbon bond formations with high diastereo- and complete enantiocontrol. Thus, this protocol includes all the advantages of domino reactions as well as asymmetric organocatalysis, and opens up a simple and flexible entry to polyfunctional cyclohexene building blocks.

METHODS

General procedure for the synthesis of cyclohexene derivatives 2. To a solution of catalyst (S)-1 (65 mg, 0.20 mmol) and nitroalkene B (1.00 mmol, 1.00 equiv.) in toluene (0.8 ml) was added subsequently under stirring aldehyde A (1.20 mmol, 1.20 equiv.) and α,β -unsaturated aldehyde C (1.05 mmol, 1.05 equiv.) at 0 °C. After 1 h the solution was allowed to reach room temperature, and stirred until complete conversion of the starting materials (16–24 h, monitored by gas chromatography). The reaction mixture was directly purified by flash column chromatography (SiO₂, ethyl acetate/pentane, 1:8–1:6) to afford the product 2. All new compounds were fully characterized (see Supplementary Information).

Characterization of 2a. The title compound 2a, (3S,4S,5R,6R)-3-methyl-5-nitro-4,6-diphenylcyclohex-1-ene carbaldehyde, was prepared according to the general procedure described above. The product was obtained as a colourless solid. The e.e. was measured by high-performance liquid chromatography (HPLC) using a chiral stationary phase (Daicel OD, *n*-heptane/*iso*-propanol 95:05) relative to the racemic sample: major isomer 13.6 min, minor isomer 17.3 min. M.p. 126 °C (recrystallized from diethyl ether); [α]_D²² (deg cm³ g^{−1} dm^{−1}) = −52.0 (*c* = 1.01 g cm^{−3}, CHCl₃). IR (KBr) 3,057, 2,815, 2,720, 2,361, 1,693, 1,650, 1,548, 1,494, 1,450, 1,365, 1,163, 761, 702 cm^{−1}. ¹H-NMR (300 MHz, CDCl₃) δ 1.23 (d, *J* = 7.2 Hz, 3H), 2.91 (dd, *J* = 10.4, 3.5 Hz, 1H), 3.45 (ddq, *J* = 10.4, 7.2, 1.4 Hz, 1H), 4.48 (s, 1H), 4.86 (s, 1H), 7.0 (m, 2H), 7.19–7.42 (m, 9H), 9.59 (s, 1H) p.p.m. ¹³C-NMR (75 MHz, CDCl₃) δ 18.7, 32.2, 43.2, 45.3, 92.4, 127.8 (2C), 128.0 (3C), 128.1, 129.1 (2C), 129.2 (2C), 136.8, 137.2, 138.8, 156.1, 192.0 p.p.m. MS (EI, 70 eV), *m/z* (%) = 321 (13) [M⁺], 274 (75), 259 (20), 245 (35), 231 (45), 215 (20), 183 (17), 167 (30), 115 (34), 105 (48), 91 (100), 77 (17). Analysis: calcd for C₂₀H₁₉NO₃ (321.37) C, 74.75; H, 5.96; N, 4.36; found C, 74.62; H, 6.22; N, 4.36.

Received 17 January; accepted 5 April 2006.

- Nicolaou, K. C., Montagnon, T. & Snyder, S. A. Tandem reactions, cascade sequences, and biomimetic strategies in total synthesis. *Chem. Commun.* 551–564 (2003).
- Wasilke, J.-C., Obrey, S. J., Baker, R. T. & Bazan, G. C. Current tandem catalysis. *Chem. Rev.* 105, 1001–1020 (2005).
- Ramón, D. J. & Yus, M. Asymmetric multicomponent reactions (AMCRs): the new frontier. *Angew. Chem. Int. Edn Engl.* 44, 1602–1634 (2005).
- Tietze, L. F. Domino reactions in organic synthesis. *Chem. Rev.* 96, 115–136 (1996).
- Tietze, L. F. & Haunert, F. in *Stimulating Concepts in Chemistry* (eds Vögtle, F., Stoddart, J. F. & Shibasaki, M.) 39–64 (Wiley-VCH, Weinheim, 2000).

6. Guo, H.-C. & Ma, J.-A. Catalytic asymmetric tandem transformations triggered by conjugate additions. *Angew. Chem. Int. Edn Engl.* **45**, 354–366 (2006).
7. Lehninger, A. L. *Principles of Biochemistry* (Worth, New York, 1993).
8. Mann, J. *Chemical Aspects of Biosynthesis* (Oxford Chemistry Primers, Oxford Univ. Press, Oxford, 1999).
9. Dalko, P. L. & Moisan, L. In the golden age of organocatalysis. *Angew. Chem. Int. Edn Engl.* **43**, 5138–5175 (2004).
10. Berkessel, A. & Gröger, H. *Asymmetric Organocatalysis* (Wiley VCH, Weinheim, 2005).
11. Seayed, J. & List, B. Asymmetric organocatalysis. *Org. Biomol. Chem.* **3**, 719–724 (2005).
12. Ramachary, D. B. & Barbas, C. F. III Towards organo-click chemistry: Development of multicomponent reactions through combination of aldol, Wittig, Knoevenagel, Michael, Diels-Alder and Huisgen cycloaddition reactions. *Chem. Eur. J.* **10**, 5323–5331 (2004).
13. Yang, J. W., Hechavarria Fonseca, M. T. & List, B. Catalytic asymmetric reductive Michael cyclization. *J. Am. Chem. Soc.* **127**, 15036–15037 (2005).
14. Huang, Y., Walji, A. M., Larsen, C. H. & MacMillan, D. W. C. Enantioselective organo-cascade catalysis. *J. Am. Chem. Soc.* **127**, 15051–15053 (2005).
15. Marigo, M., Schulte, T., Franzén, J. & Jørgensen, K. A. Asymmetric multicomponent domino reactions and highly enantioselective conjugated addition of thiols to α,β -unsaturated aldehydes. *J. Am. Chem. Soc.* **127**, 15710–15711 (2005).
16. Hayashi, Y., Gotoh, H., Hayashi, T. & Shoji, M. Diphenylprolinol silyl ethers as efficient organocatalysts for the asymmetric Michael reaction of aldehydes and nitroalkenes. *Angew. Chem. Int. Edn Engl.* **44**, 4212–4215 (2005).
17. Prieto, A., Halland, N. & Jørgensen, K. A. Novel imidazolidine-tetrazole organocatalyst for asymmetric conjugate addition to nitroalkanes. *Org. Lett.* **7**, 3897–3900 (2005).
18. Berner, O. M., Tedeschi, L. & Enders, D. Asymmetric Michael-additions to nitroalkenes. *Eur. J. Org. Chem.* 1877–1894 (2002).
19. Enders, D., Kipphardt, H., Gerdes, P., Breña-Valle, L. J. & Bhushan, V. Large scale preparation of versatile chiral auxiliaries derived from (S)-proline. *Bull. Soc. Chim. Belg.* **97**, 691–704 (1988).
20. Franzén, J. et al. A general organocatalyst for direct α -functionalization of aldehydes: stereoselective C–C, C–N, C–F, C–Br, and C–S bond-forming reactions. Scope and mechanistic insights. *J. Am. Chem. Soc.* **127**, 18296–18304 (2005).
21. List, B., Pojarliev, P. & Martin, H. J. Efficient proline-catalyzed Michael additions of unmodified ketones to nitro olefins. *Org. Lett.* **3**, 2423–2425 (2001).
22. Enders, D. & Seki, A. Proline-catalyzed enantioselective Michael additions of ketones to nitrostyrene. *Synlett*, 26–28 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A figure summarizing the main result of this paper is included.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft (SPP Organokatalyse) and the Fonds der Chemischen Industrie (thanked by C.G. for a Kekulé fellowship). We thank BASF AG, Degussa AG, Bayer AG and Wacker-Chemie for the donation of chemicals, and J. Runkel for the NOE measurements.

Author Contributions M.R.M.H. and C.G. contributed equally to this work. The X-ray structure analysis was performed by G.R.

Author Information Crystallographic data have been deposited with the Cambridge Crystallographic Data Centre, accession number CCDC 295450, and are available via www.ccdc.cam.ac.uk/data_request/cif. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.E. (Enders@RWTH-Aachen.de).

LETTERS

The importance of the diurnal and annual cycle of air traffic for contrail radiative forcing

Nicola Stuber¹, Piers Forster^{1†}, Gaby Rädcl¹ & Keith Shine¹

Air traffic condensation trails, or contrails, are believed to have a net atmospheric warming effect¹, although one that is currently small compared to that induced by other sources of human emissions. However, the comparably large growth rate of air traffic requires an improved understanding of the resulting impact of aircraft radiative forcing on climate². Contrails have an effect on the Earth's energy balance similar to that of high thin ice clouds³. Their trapping of outgoing longwave radiation emitted by the Earth and atmosphere (positive radiative forcing) is partly compensated by their reflection of incoming solar radiation (negative radiative forcing). On average, the longwave effect dominates and the net contrail radiative forcing is believed to be positive^{1,2,4}. Over daily and annual timescales, varying levels of air traffic, meteorological conditions, and solar insolation influence the net forcing effect of contrails. Here we determine the factors most important for contrail climate forcing using a sophisticated radiative transfer model^{5,6} for a site in southeast England, located in the entrance to the North Atlantic flight corridor. We find that night-time flights during winter (December to February) are responsible for most of the contrail radiative forcing. Night flights account for only 25 per cent of daily air traffic, but contribute 60 to 80 per cent of the contrail forcing. Further, winter flights account for only 22 per cent of annual air traffic, but contribute half of the annual mean forcing. These results suggest that flight rescheduling could help to minimize the climate impact of aviation.

Contrails form in the wake of aircraft only when the surrounding atmospheric conditions—in connection with the characteristics of the aircraft exhaust—are favourable⁷. In ice-supersaturated regions⁸ contrails can exist for several hours. Flight management systems could be modified to reduce contrail radiative forcing. One way to reduce contrail coverage and forcing is by changing flight routes and/or cruising altitudes to avoid ice-supersaturated regions^{9,10}. Another is to reduce the radiative forcing of contrails when they are present. Because of the compensation between positive longwave and negative shortwave forcings, shifting air traffic to times when the negative effects are largest would reduce the net contrail radiative forcing¹¹. Our work aims to understand which parts of the diurnal and annual cycle of relevant parameters have the largest impact on the net radiative forcing of contrails.

We used AERO2k flightdata¹² over Herstmonceux (southeast England; longitude 0.32° E, latitude 50.90° N), a location for which vertical temperature and humidity profiles are available from radiosonde observations. AERO2k holds information on height-resolved distances flown in a 1° × 1° grid for each month in 2002 and for four six-hourly time periods—starting at midnight Greenwich Mean Time (GMT)—averaged over one week in June 2002. Using the size of the gridbox and the distance travelled in it, we first calculated the maximum possible contrail cover over Herstmonceux, initially assuming that all aircraft produce persistent contrails. In accordance with measurements¹³ and model simulations¹⁴ of line-shaped, persistent

contrails, we assumed a contrail width of 2 km and a contrail lifetime of 2 hours. We scaled the data obtained using the monthly total air traffic over Herstmonceux to get diurnally resolved data for each month and assumed that these data for the year 2002 are also representative for the year 2003, for which Met Office radiosonde profiles were available. We corrected the systematic dry bias in the radiosonde humidity measurements¹⁵, and applied the Schmidt–Appleman thermodynamic criterion for contrail formation^{16,17} to determine the contrail cover. For contrail formation the mixture of aircraft exhaust and ambient air must reach water saturation; contrail persistency requires ice supersaturation.

We compared predictions of contrail-favourable conditions made on the basis of radiosonde data with our observations of the occurrence of persistent contrails. Data from the radiosonde ascent at Herstmonceux successfully predicted whether or not persistent contrails formed over Reading in 60 of the 81 cases analysed (see Table 1). Statistical tests of the resulting contingency table showed that these predictions were not due to chance (see the uncertainty analysis in Methods section ‘Statistical significance of radiosonde contrail predictions’ and Table 1).

Air traffic over Herstmonceux shows a distinct annual cycle (Fig. 1a) with a summer (June, July, August; JJA) maximum and a winter (December, January, February; DJF) minimum, with approximately 20% fewer flights. Flight restrictions at nighttime mean that approximately 75% of the total air traffic over Herstmonceux occurs between 06:00 and 18:00 GMT. Figure 1b shows the contrail frequency of occurrence, that is, the percentage of days each month when persistent contrails could occur. In mid-latitudes the upper tropospheric relative humidity has its lowest values during summer¹⁸, so the contrail frequency is lowest in summer. The radiosonde data indicate that contrail formation over Herstmonceux is almost twice as likely in winter than in summer. Figure 1 shows that the peaks of air traffic and contrail-favourable conditions occur in different seasons.

Using the AERO2k data and only introducing a contrail if atmospheric conditions were favourable at the actual cruising altitudes, we calculated the instantaneous radiative forcing at the top of

Table 1 | Predicted and observed contrails over Herstmonceux

	Persistent contrails predicted	Either short-lived contrails or no contrails predicted
Observations of persistent contrails	24	18
No persistent contrails observed	3	36

Contingencies are shown for persistent contrails observed over Reading, and predicted from radiosonde data for Herstmonceux. Observations were made at Reading between July 2004 and June 2005 at least four times a day and compared to persistent contrail-favourable conditions predicted on the basis of radiosonde ascents made at Herstmonceux within about one hour of the observation at Reading.

¹Department of Meteorology, The University of Reading, Reading RG6 6BB, UK. [†]Present address: School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK.

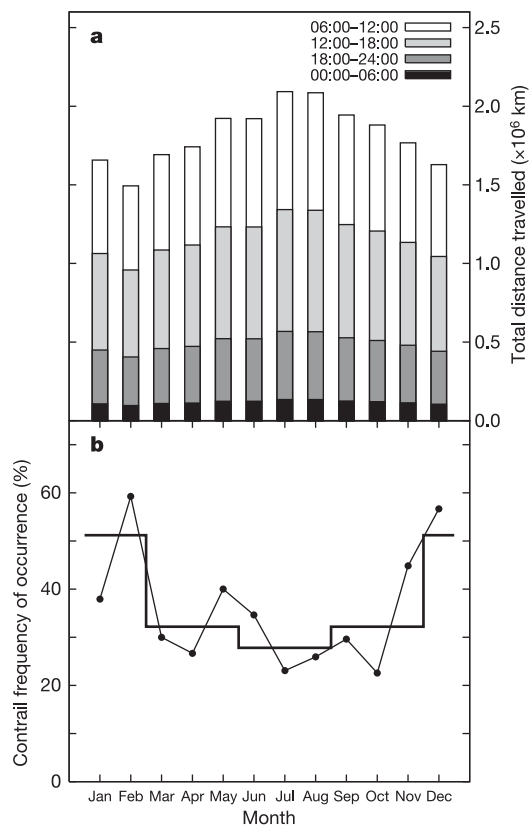


Figure 1 | Air traffic and contrail occurrence over Herstmonceux. **a**, Annual cycle of total column air traffic over Herstmonceux. Contributions of the four individual six-hour time periods are indicated by differently shaded bars. **b**, Annual cycle of the frequency of possible persistent contrail occurrence over Herstmonceux; that is, the number of days each month with persistent contrail-favourable conditions at one or more levels. This frequency is based solely on the meteorological conditions, indicating the potential percentage of contrail days. It does not take into account the actual vertical profile of air traffic. The heavy black line shows the seasonal mean.

the atmosphere due to contrails over Herstmonceux, using a sophisticated radiative transfer model^{5,6}. We allowed for the full diurnal cycle of shortwave insolation but used a single radiosonde atmospheric profile at all times of day—the choice of a single background profile was found to have a negligible impact on the radiative forcing. We complemented the radiosonde data with climatological profiles providing information on ozone, higher level temperatures and humidities, and surface albedo. We assumed a contrail visible optical depth of 0.1, standard-sized contrail particles¹⁹ and random overlap for contrails at different altitudes. As the radiosonde data gives no information on natural clouds, the calculations were performed for otherwise clear-sky conditions. However, we investigated the effect of this assumption through sensitivity studies (see the uncertainty analysis in Methods section ‘Effect of natural clouds on contrail radiative forcing’).

The diurnal, annual mean radiative forcings (Fig. 2) are 0.78 and -0.54 W m^{-2} in the longwave and shortwave, respectively, resulting in a net forcing of 0.23 W m^{-2} . The locally large net forcing is due to the location of Herstmonceux in the entrance to the North Atlantic flight corridor. The diurnal variation in longwave radiative forcing reflects the daily variation in air traffic (Fig. 1a). Each of the two time periods 06:00–12:00 GMT and 12:00–18:00 GMT contributes roughly 50% to the diurnal, annual mean shortwave radiative forcing. Strikingly, because of the near cancellation between shortwave and longwave forcings for these time periods, almost 82% of the annual mean net radiative forcing is due to flights between 18:00 and

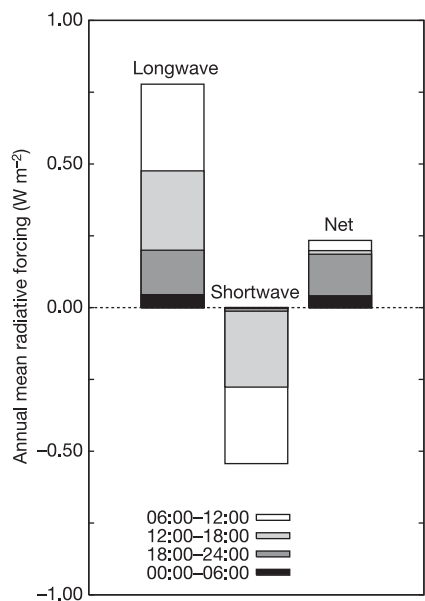


Figure 2 | Annual mean longwave, shortwave, and net radiative forcing due to persistent contrails over Herstmonceux. The contributions of flights occurring during different time periods to the diurnal mean values are indicated by differently shaded bars.

06:00 GMT, even though these night flights are responsible for only 25% of the air traffic.

The annual cycle of forcings (Fig. 3) reveals the complex interaction and competition of different effects: Time-averaged shortwave forcings are affected by daylength and solar zenith angle. With all other parameters held fixed, the shortwave effect will increase with solar zenith angle and the duration of sunlight. The effect of these two factors is superimposed on the annual variation in meteorological conditions and air traffic. Daytime shortwave forcings are smallest in summer because the combination of low solar zenith angles and smaller chances of forming a contrail dominates the effect of comparably higher summertime air traffic and longer daytime hours.

The annual cycle of longwave radiative forcing shows the dominant role of variations in meteorological conditions and hence contrail frequency of occurrence, compared to variations in air traffic. The cancellation between longwave and shortwave forcings is large throughout the year and the net forcing can sometimes be negative (Fig. 3b). The seasonal cycle of the diurnal mean radiative forcing (heavy black line in Fig. 3b) emphasizes the importance of winter (DJF) flights, which contribute 50% to the annual mean contrail radiative forcing despite being responsible for only 22% of the flights.

The net forcing is the residual from subtracting two comparatively large numbers, so its magnitude and sign are very sensitive to uncertainties in input parameters. Important sources of uncertainty are the contrail’s ice crystal size and optical depth (see the uncertainty analysis in Methods section ‘Sensitivity of radiative forcing to

Table 2 | Percentage contribution of night flights to radiative forcing over Herstmonceux

	Relative size 0.5	Relative size 1.0	Relative size 2.0
Optical depth 0.05	116	81	63
Optical depth 0.1	113	82	62
Optical depth 0.3	115	81	64
Optical depth 0.5	123	87	67

The percentage contribution of flights between 18:00 and 06:00 to the diurnal, annual mean net radiative forcing over Herstmonceux is shown. Percentages are given for different contrail particle sizes and contrail visible optical depths.

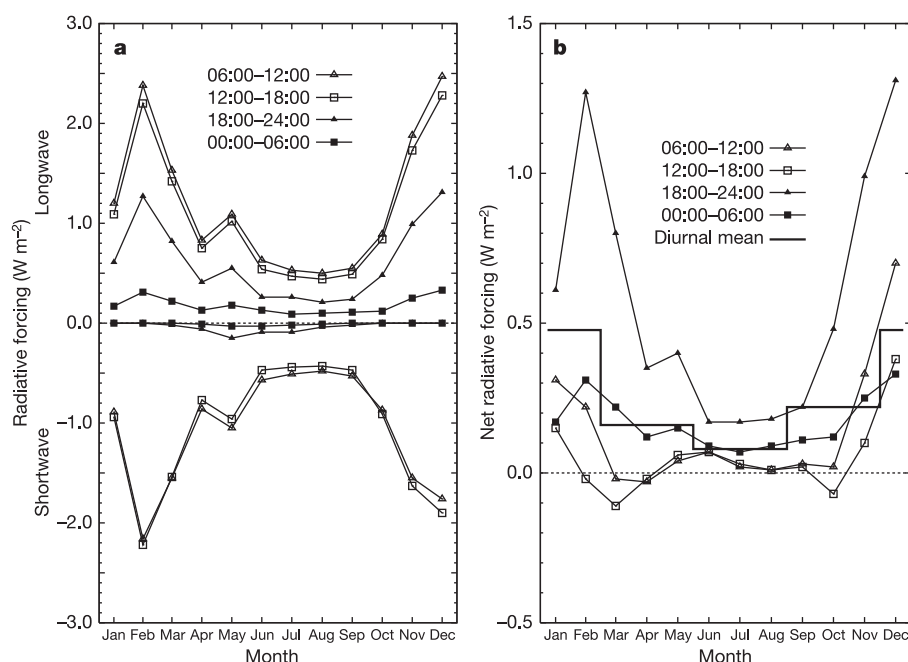


Figure 3 | Annual cycle of the diurnal-mean contrail radiative forcing over Herstmonceux for the four time periods. a, Longwave (positive) and shortwave (negative) radiative forcing. b, Net radiative forcing. The heavy black line shows the seasonally averaged diurnal-mean radiative forcing.

contrail optical depth and particle size' and Table 2). We found that the contribution of night-time flights to the total forcing decreased when crystal sizes were increased. However, even for the largest ice crystals, night-time flights were still found to contribute at least 62% to the daily mean forcing, depending on the contrail optical depth. Decreasing particle sizes implies a decrease in net forcing. However, this decrease in absolute forcing is accompanied by a further increase in the relative importance of night-time flights, as daytime net forcings become negative.

Natural clouds were omitted in the calculations described above because cloud data was not available from the radiosondes. In sensitivity calculations using climatological cloud distributions (see the uncertainty analysis in Methods section 'Effect of natural clouds on contrail radiative forcing') we found that natural clouds reduce both the shortwave and longwave contrail radiative forcing. However, their effect on the net forcing is variable, increasing daytime net forcings and decreasing night-time net forcings. As a consequence, the contribution of night-time flights to the diurnal mean net radiative forcing was smaller for cloudy (59%) than for clear-sky (82%) conditions. However, given their comparably small share of daily total air traffic (25%), night-time flights still contribute disproportionately to the diurnal mean forcing.

The sensitivity studies showed that changing contrail optical properties or the presence of natural clouds did not have an appreciable effect on the relative importance of the winter season for the annual mean contrail radiative forcing.

In summary, flights between 18:00 and 06:00 GMT have a disproportionate effect on the daily mean radiative forcing over Herstmonceux. Given the current flight schedules, these flights account for only 25% of total air traffic, yet they contribute 60 to 80% to the net radiative forcing. Flights during winter (DJF) are almost twice as likely to form a contrail as are summer (JJA) flights and contribute 50% to the annual mean radiative forcing.

Emissions from aircraft affect climate in numerous ways^{1,2}. However, the radiative properties of contrails and their short lifetime make them the only mechanism for which a rescheduling of flight times can significantly change the radiative forcing. Our results show that, in terms of radiative forcing, the southeast of England already benefits from night-flying restrictions. At locations without such

restrictions the contribution of night-time flights to daily mean contrail radiative forcing could be even larger than at Herstmonceux. For example, for South-East Asia, the AERO2k data set indicates that 40% of flights occur during night, and we find that these flights contribute 73% to the diurnal, annual mean forcing (see the uncertainty analysis in Methods section 'Effect of natural clouds on contrail radiative forcing'). In view of this, we argue that shifting air traffic from night-time to daytime would help to minimize the climate effect of contrails.

METHODS

Statistical significance of radiosonde contrail predictions. We predicted persistent contrail-favourable conditions by applying the Schmidt–Appleman thermodynamic criterion to the radiosonde data. An error in the prediction of whether a contrail could occur might be introduced either by errors in the threshold values of temperature and humidity given by the Schmidt–Appleman criterion and/or by errors in the radiosonde data. False decisions would have a direct impact on the contrail cover and hence on the radiative forcing.

The Schmidt–Appleman criterion has been tested and verified in experimental studies^{20,21}. We therefore assume that any false decisions are due solely to errors in the radiosonde data. We compared predictions of contrail-favourable conditions made on the basis of Herstmonceux radiosonde data with observations of persistent contrails over Reading. The distance between Reading and Herstmonceux is smaller than the typical spatial scale of contrail clusters^{22,23} and the mean extension of ice-supersaturated regions^{24,25}. We statistically tested the resulting contingency table (Table 1) using different tests. The 'odds ratio'²⁶ compares the odds of making a true forecast with those of making a false one. It is 1 if there is no correlation between observations and predictions, and larger or smaller than 1 for a positive or negative correlation. For our case we get an odds ratio of 16. The probability that there is a positive correlation between observing persistent contrails over Reading and predicting contrails from Herstmonceux radiosonde ascents exceeds 99.5%.

In addition, we calculated the significance of the Peirce skill score, which gives a measure of accuracy for both the 'yes' and 'no' events and compares the hit rate with the false alarm rate. The score takes values between -1 and 1 , being zero in case of a random forecast and greater than zero if there is some forecasting skill. We calculated a score of 0.49 .

Sensitivity of radiative forcing to contrail optical depth and particle size. In view of the uncertainties associated with contrail optical properties, in addition to the reference experiment described above we conducted a number of sensitivity experiments in which we changed both the contrail particle size and the contrail optical depth. Increasing the particle size results in an increase in

the magnitude of the shortwave forcing, and vice versa, while the longwave forcing is practically unaffected. In contrast, with increasing optical depths the magnitude of both the shortwave and the longwave radiative forcing increases.

The IPCC's estimate of contrail radiative forcing¹ relied on an optical depth of 0.3. However, as typical values are now believed to be lower^{27,28} we adopted a value of 0.1 in our reference experiment, decreasing it to 0.05 and increasing it to 0.5 in the sensitivity experiments. Additionally, we varied the contrail particle size. The standard size used in the reference experiment is based on contrail ice-crystal size distributions derived from both *in situ* measurements and a temperature-dependent parameterisation¹⁹. We conducted experiments with relative contrail particle sizes of 0.5 and 2.0. These are equivalent to halving or doubling the standard width as well as the standard length of the hexagonal ice particles, respectively. Table 2 gives the percentage of diurnal mean contrail radiative forcing that is due to flights between 18:00 and 06:00 GMT for different contrail optical depths and particle sizes.

Effect of natural clouds on contrail radiative forcing. It has been shown¹¹ that the presence of natural clouds has a negligible effect on the global mean, annual and diurnal mean net contrail radiative forcing. However, the study did not give details on the effects of clouds on the diurnal and annual cycle of contrail radiative forcing.

To determine the effects of natural cloudiness on contrail radiative forcing we combined AERO2k data with analysis data from the integrated forecast system of the European Centre for Medium-Range Weather Forecasts to calculate global contrail cover and calibrated the contrail cover to an annual and diurnal area mean value of 0.375% (ref. 29) in the Bakan area²². As input for the radiative transfer model we derived atmospheric profiles using a three-dimensional climatology compiled at the University of Reading. This climatology is based on satellite, aircraft and ground-based observations and provides long-term monthly mean profiles of temperatures and the mixing ratios of water vapour and ozone extending up to 1 hPa. Information is also given about the surface albedo and the amount, optical depth and height of low, mid- and high-level clouds. Cloud information is based on ISCCP C2 data³⁰.

We analysed the results for selected locations. For western Europe (Herstmonceux), clouds reduce the contribution of night-time flights to annual, diurnal mean contrail radiative forcing to 59% (from 82% for cloud-free conditions). For the East Coast of the United States, 36% of flights occur during the night, contributing 53% to the diurnal, annual mean radiative forcing. For South-East Asia night flights (40%) contribute 73%, and in the North Atlantic flight corridor (48%) they contribute 58%.

Received 3 August 2005; accepted 5 May 2006.

1. The IPCC Working Group Aviation and the Global Atmosphere—A Special Report of IPCC Working Groups I and III (eds Penner, J. E. et al.) (Cambridge Univ. Press, Cambridge, UK, 1999).
2. Sausen, R. et al. Aviation radiative forcing in 2000: An update on the IPCC (1999). *Meteorol. Z.* **14**, 555–561 (2005).
3. Hartmann, D. L., Ockert-Bell, M. E. & Michelsen, M. L. The effect of cloud type on Earth's energy balance: global analysis. *J. Clim.* **5**, 1281–1304 (1992).
4. Minnis, P., Schumann, U., Doelling, D. R., Gierens, K. M. & Fahey, D. W. Global distribution of contrail radiative forcing. *Geophys. Res. Lett.* **26**, 1853–1856 (1999).
5. Fu, Q. & Liou, K. N. On the correlated k-distribution method for radiative transfer in nonhomogeneous atmospheres. *J. Atmos. Sci.* **49**, 2139–2156 (1992).
6. Fu, Q. & Liou, K. N. Parameterization of the radiative properties of cirrus clouds. *J. Atmos. Sci.* **50**, 2008–2025 (1993).
7. Schumann, U. On conditions for contrail formation from aircraft exhausts. *Meteorol. Z.* **5**, 4–23 (1996).
8. Gierens, K., Schumann, U., Helten, M., Smit, H. G. J. & Marengo, A. A distribution law for relative humidity in the upper troposphere and lower stratosphere derived from three years of MOZAIC measurements. *Ann. Geophys.* **17**, 1218–1226 (1999).
9. Sausen, R., Nodorp, D., Land, C. & Deidewig, F. *Ermittlung optimaler Flughöhen und Flugrouten unter dem Aspekt minimaler Klimawirksamkeit* 96–13 (DLR-Forschungsbericht, Cologne, Germany, 1996).
10. Williams, V. & Noland, R. B. Variability of contrail formation conditions and the implications for policies to reduce the climate impacts of aviation. *Transp. Res. Part D* **10**, 269–280 (2005).
11. Myhre, G. & Stordal, F. On the tradeoff of the solar and thermal infrared radiative impact of contrails. *Geophys. Res. Lett.* **28**, 3119–3122 (2001).
12. Eysers, C. J. et al. *AERO2k Global Aviation Emissions Inventories for 2002 and 2025* (QinetiQ Ltd, Farnborough, UK, 2004).
13. Schumann, U. On the effect of emissions from aircraft engines on the state of the atmosphere. *Ann. Geophys.* **12**, 365–384 (1994).
14. Gierens, K. M. Numerical simulations of persistent contrails. *J. Atmos. Sci.* **53**, 3333–3348 (1996).
15. Wang, J., Cole, H. L., Carlson, D. J., Miller, E. R. & Beierle, K. Corrections of humidity measurement errors from the Vaisala RS80 radiosonde—Application to TOGA COARE Data. *J. Atmos. Oceanic Technol.* **19**, 981–1002 (2002).
16. Schmidt, E. Die Entstehung von Eisnebel aus den Auspuffgasen von Flugmotoren. *Schr. deutsch. Akad. Luftfahrtforsch.* **44**, 1–15 (1941).
17. Appleman, H. The formation of exhaust condensation trails by jet aircraft. *Bull. Am. Meteorol. Soc.* **34**, 14–20 (1953).
18. Kley, D. et al. (eds) *SPARC Assessment of Upper Tropospheric and Stratospheric Water Vapour WCRP-T13*, WMO/TD-No. 1043 (SPARC, Verrières le Buisson Cedex, 2000).
19. Strauss, B., Meerkötter, R., Wissinger, B., Wendling, P. & Hess, M. On the regional climatic impact of contrails: microphysical and radiative properties of contrails and natural cirrus clouds. *Ann. Geophys.* **15**, 1457–1467 (1997).
20. Jensen, E. J. et al. Environmental conditions required for contrail formation and persistence. *J. Geophys. Res.* **103** (D4), 3929–3936 (1998).
21. Schumann, U. Influence of propulsion efficiency on contrail formation. *Aerosp. Sci. Technol.* **4**, 391–401 (2000).
22. Bakan, S., Betancor, M., Gayler, V. & Grassl, H. Contrail frequency over Europe from NOAA-satellite images. *Ann. Geophys.* **12**, 962–968 (1994).
23. Mannstein, H., Meyer, R. & Wendling, P. Operational detection of contrails from NOAA-AVHRR-data. *Int. J. Remote Sens.* **20**, 1641–1660 (1999).
24. Detwiler, A. & Pratt, R. Clear-air seeding: opportunities and strategies. *J. Weath. Modif.* **16**, 46–60 (1984).
25. Gierens, K. & Spichtinger, P. On the size distribution of ice-supersaturated regions in the upper troposphere and lowermost stratosphere. *Ann. Geophys.* **18**, 499–504 (2000).
26. Stephenson, D. B. Use of the “odds ratio” for diagnosing forecast skill. *Weath. Forecast.* **15**, 221–232 (2000).
27. Meyer, B., Mannstein, H., Meerkötter, R., Schumann, U. & Wendling, P. Regional radiative forcing by line-shaped contrails derived from satellite data. *J. Geophys. Res.* **107** (D10), 4104, doi: 10.1029/2001JD000426 (2002).
28. Ponater, M., Marquart, S. & Sausen, R. Contrails in a comprehensive climate model: Parameterization and radiative forcing results. *J. Geophys. Res.* **107** (D13), 4164, doi: 10.1029/2001JD000429 (2002).
29. Marquart, S., Ponater, M. & Sausen, R. Future development of contrail cover, optical depth, and radiative forcing: Impacts of increasing air traffic and climate change. *J. Clim.* **16**, 2890–2904 (2003).
30. Rossow, W. B., Garder, L., Lu, P.-J. & Walker, A. *International Satellite Cloud Climatology Project (ISCCP) Documentation of Cloud Data WMO/TD 266* (World Climate Research Programme, Geneva, 1988).

Acknowledgements We thank Q. Fu for providing the basic radiative transfer code and C. Eysers for the AERO2k data. Radiosonde data was provided by the British Atmospheric Data Centre (BADC). A. Tompkins (ECMWF) provided us with analysis data. N.S. was supported by the Department for Transport, and G.R. by the Department of Trade and Industry, and Airbus.

Author Contributions N.S. was the principal researcher. G.R. performed analysis of contrail observations. P.F. led the research with significant contributions from K.S.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.S. (n.stuber@reading.ac.uk).

LETTERS

Speciation by hybridization in *Heliconius* butterfliesJesús Mavárez^{1*}, Camilo A. Salazar^{2*}, Eldredge Bermingham¹, Christian Salcedo², Chris D. Jiggins³ & Mauricio Linares²

Speciation is generally regarded to result from the splitting of a single lineage. An alternative is hybrid speciation, considered to be extremely rare, in which two distinct lineages contribute genes to a daughter species. Here we show that a hybrid trait in an animal species can directly cause reproductive isolation. The butterfly species *Heliconius heurippa* is known to have an intermediate morphology and a hybrid genome¹, and we have recreated its intermediate wing colour and pattern through laboratory crosses between *H. melpomene*, *H. cydno* and their F₁ hybrids. We then used mate preference experiments to show that the phenotype of *H. heurippa* reproductively isolates it from both parental species. There is strong assortative mating between all three species, and in *H. heurippa* the wing pattern and colour elements derived from *H. melpomene* and *H. cydno* are both critical for mate recognition by males.

Homoploid hybrid speciation—hybridization without change in chromosome number—is considered very rare^{2–4}. This has been explained by the theoretical prediction that reproductive isolation between hybrids and their parents is difficult to achieve^{3,5,6}. However, if a hybrid phenotype directly causes reproductive isolation from parental taxa, this difficulty can be overcome. Such a role for a hybrid phenotype has been convincingly demonstrated only in *Helianthus* sunflowers⁷. In animals, the evidence for homoploid hybrid speciation is less convincing. Putative hybrid species are known with mixed genomes^{8–11}, but in these examples shared genetic variation could also be a result of introgression subsequent to a bifurcating speciation event.

Heliconius cydno and *H. melpomene* are two closely related species that overlap extensively in lower Mesoamerica and the Andes¹². Speciation in these butterflies has not involved any change in chromosome number¹³ but is instead associated with shifts in colour patterns that generate both assortative mating and postzygotic isolation due to predator-mediated selection^{14–17}. *Heliconius cydno* is black with white and yellow marks, whereas *H. melpomene* is black with red, yellow and orange marks. Both species exhibit strong positive assortative mating based on their wing colour patterns and also differ in habitat use¹⁸ and host plant preference¹⁹, but inter-specific hybrids do occur at low frequency in the wild¹⁵. *Heliconius heurippa* has an intermediate wing pattern, which has led to the suggestion that this is a hybrid species^{1,20}. Its hindwing is indistinguishable from that of sympatric *H. m. melpomene*, whereas the yellow band on its forewing is similar to that of parapatric *H. cydno cordula*. Ecologically, *H. heurippa* is most similar to *H. cydno*, which it replaces geographically in the eastern Andes of Colombia.

Here we first establish that *H. heurippa* is currently genetically isolated from its putative parents and provide evidence that its genome is of hybrid origin. A bayesian assignment analysis using 12 microsatellite loci scored in populations from Panama, Colombia and Venezuela divides *H. cydno* ($n = 175$), *H. melpomene* ($n = 167$)

and *H. heurippa* ($n = 46$) individuals into three distinct clusters (Fig. 1). Hence, *H. heurippa* is genetically more differentiated than any geographic race sampled of either species. Moreover, analyses of polymorphism at two nuclear genes (*Invected* and *Distal-less*) show no allele sharing between *H. cydno* and *H. melpomene*, whereas the *H. heurippa* genome appears as an admixture, sharing allelic variation with both putative parental species (Supplementary Fig. 2, and C.S., C.D.J. and M.L., unpublished observations).

To test the hypothesis of a hybrid origin for the *H. heurippa* colour pattern, we performed inter-specific crosses between *H. cydno*

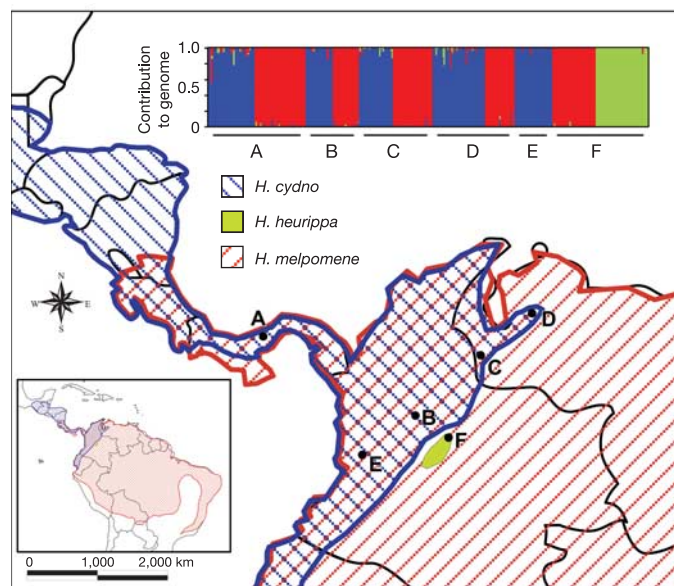
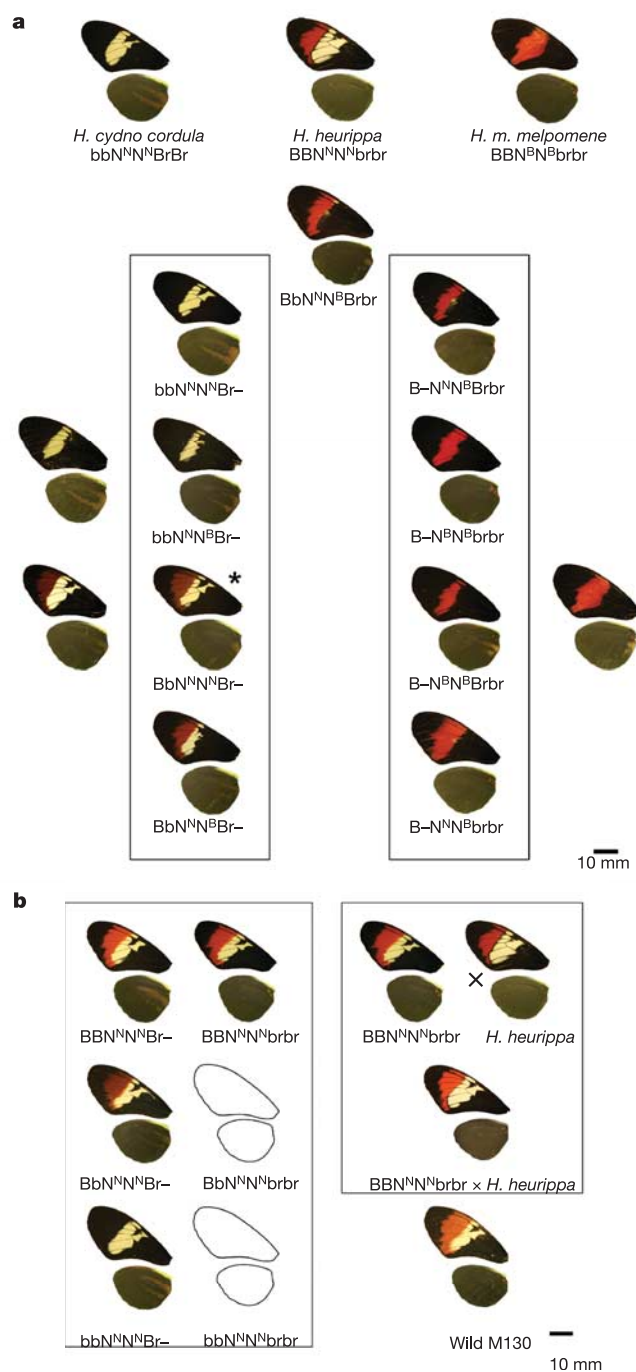


Figure 1 | Geographic distributions and genetic differentiation between *H. cydno*, *H. melpomene* and *H. heurippa*. *H. heurippa* is sympatric with *H. melpomene* and parapatric with *H. cydno* in eastern Colombia. We used the software Structure 2.1 (ref. 29) with the multilocus microsatellite data set to assign individuals to species and detect admixed individuals (namely hybrids). We ran Structure 2.1, varying the burn-in (10^4 to 10^5) and run length (10^5 to 10^6), number of clusters (one to four), ancestry type (with and without admixture) and allele frequency estimation (correlated and independent) to obtain the highest probability model for the data set. The upper inset shows the results obtained with the best model (three clusters, admixture and independent estimations of allele frequencies). The relative contributions of the three clusters to each individual's genome are shown in the following colours: blue, *H. cydno*; red, *H. melpomene*; green, *H. heurippa*. Collection site codes: A, Pipeline Road, Panama; B, Parcela 33, Colombia; C, San Cristóbal, Venezuela; D, La Gira, Venezuela; E, Ocache, Colombia; F, Villavicencio, Colombia. An expanded view of the clusters is shown in Supplementary Fig. 1.

¹Smithsonian Tropical Research Institute, Apartado postal 0843-03092, Panamá, República de Panamá. ²Instituto de Genética, Universidad de los Andes, Carrera 1E No 18A-10, PO Box 4976, Santafé de Bogotá D.C., Colombia. ³Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK.

*These authors contributed equally to this work.



cordula and *H. m. melpomene* to reconstruct the steps of introgressive hybridization that could have given rise to *H. heurippa*. The colour pattern differences between *H. m. melpomene* and *H. cydno cordula* are determined largely by three co-dominant loci controlling the red and yellow bands on the forewing and the brown pincer-shaped mark on the ventral hindwing (see Fig. 2a)^{21,22}. Most *H. cydno* × *H. melpomene* F₁ hybrids seem intermediate to both parents (Fig. 2a), with both a yellow (*cydno*) and a red (*melpomene*) band in the median section of the forewing, whereas the ventral side of the hindwing shows a reduced brown mark intermediate between the parental species.

Female F₁ hybrids resulting from crosses between *H. melpomene* and *H. cydno* are sterile in accordance with Haldane's rule^{1,23}, and thus only male F₁ hybrids backcrossed to either *H. cydno cordula* or *H. m. melpomene* females resulted in offspring. Backcrosses to *H. melpomene* produced offspring very similar to pure *H. m. melpomene*, and further backcross generations never produced

Figure 2 | Reconstruction of the *H. heurippa* wing pattern. All fore- and hind-wings shown in dorsal and ventral views, respectively. **a**, First row, *H. cydno cordula*, *H. heurippa* and *H. m. melpomene*; second row, *H. cydno cordula* × *H. m. melpomene* F₁ hybrid. Backcrosses to *H. cydno cordula* and *H. m. melpomene* are shown in the left and right boxes, respectively. Other individuals are wild hybrids from San Cristóbal, Venezuela, shown next to their putative genotypes. Three major loci regulate the patterns: complete (BB, *H. melpomene*), intermediate (Bb, heterozygotes) or no (bb, *H. cydno*) expression of a red band on the dorsal forewing; complete (N^NN^N, *H. cydno*), intermediate (N^NN^B, heterozygotes) or no (N^BN^B, *H. melpomene*) expression of a yellow band on the dorsal forewing; and complete (BrBr, *H. cydno*), intermediate (Brbr, heterozygotes) or no (brbr, *H. melpomene*) expression of a brown pincer-shaped mark on the ventral hindwing. Further crosses were performed using individuals with the phenotype marked with an asterisk (see **b**). **b**, Left box: offspring from crosses between individuals marked with an asterisk in **a**. The B and Br loci are linked, explaining the absence of the two recombinant genotypes BbN^NN^Nbrbr and bbN^NN^Nbrbr. Right box: offspring of a cross between a laboratory hybrid with genotype BBN^NN^Bbrbr and *H. heurippa*, showing that the pattern breeds true. The other individual (M130) is a wild hybrid from San Cristóbal, Venezuela, with a phenotype very similar to *H. heurippa*.

individuals with forewing phenotypes similar to *H. heurippa* (Fig. 2a). However, after only two generations a phenotype virtually identical to *H. heurippa* (Supplementary Fig. 3) was produced by backcrossing an F₁ male to an *H. cydno cordula* female and then mating selected offspring of this cross (Fig. 2b). In offspring of crosses between these *H. heurippa*-like individuals the pattern breeds true, showing that they are homozygous for the red forewing band (BB) and the absence of brown hindwing marks (brbr) characteristic of *H. melpomene*, and similarly homozygous for the yellow forewing band (N^NN^N) derived from *H. cydno*. The pattern of these *H. heurippa*-like individuals also breeds true when crossed to wild *H. heurippa* (Fig. 2b), implying that pattern genes segregating in our crosses are homologous with those in wild *H. heurippa*.

Furthermore, in a wild population of sympatric *H. m. melpomene* and *H. cydno cordula* in San Cristóbal, Venezuela, we observed natural hybrids at an unusually high frequency (8%), including some individuals very similar to our laboratory-produced *H. heurippa*-like butterflies (Fig. 2b). Microsatellite data show that these individuals have genotypes indistinguishable from that of *H. cydno* and must therefore be at least fifth-generation backcrosses (Supplementary Fig. 4). This shows that multiple generations of backcrossing can occur in the wild and that female hybrid sterility is not a complete barrier to introgressive hybridization. The fact that the *H. heurippa* pattern can be generated by laboratory crosses between *H. melpomene* and *H. cydno*, and is also observed in wild hybrids between the two species, establishes a probable natural route for the hybrid origin of *H. heurippa*.

The next step in species formation is reproductive isolation. We therefore tested the degree to which *H. heurippa* is isolated from *H. melpomene* and *H. cydno* by assortative mating. No-choice mating experiments showed a reduced probability of mating in all inter-specific comparisons, with *H. heurippa* females particularly unlikely to mate with either *H. cydno* or *H. melpomene* (Table 1). When a male of each species was presented with a single female, *H. heurippa* males were tenfold more likely to court their own females than the other species (Supplementary Fig. 5). In mating experiments with choice, there was similarly strong assortative mating, although occasional matings between *H. cydno* and *H. heurippa* were observed (Table 2). Isolation due to assortative mating, on average more than 90% between *H. heurippa* and *H. melpomene* and more than 75% between *H. heurippa* and *H. cydno*, is therefore considerably greater than that caused by hybrid sterility (about 25% isolation between *H. heurippa* and *H. melpomene*, and zero between *H. heurippa* and *H. cydno*)¹ or predator selection against hybrids (about 50%)²⁴. Therefore, strong assortative mating, in combination with geographic isolation

Table 1 | Relative mating probabilities in no-choice experiments

Male	<i>H. melpomene</i>	Female <i>H. cydno</i>	<i>H. heurippa</i>
<i>H. melpomene</i>	1 (17)	0.178 (0.084–0.309, 45)	0.073 (0.023–0.163, 55)
<i>H. cydno</i>	0.120 (0.048–0.231, 50)	1 (27)	0.022 (0.001–0.096, 45)
<i>H. heurippa</i>	0.100 (0.031–0.022, 40)	0.440 (0.255–0.637, 44)	1 (22)

For each female type, probabilities were estimated relative to that of intra-specific mating, which was set to 1. Numbers in parenthesis show the 95% maximum-likelihood support limits and the number of females used.

from *H. cydno* and postzygotic isolation from *H. melpomene* has contributed to the speciation of *H. heurippa*.

We next investigated the role of colour pattern in mate choice. Experiments with dissected wings showed that both elements of the forewing colour pattern of *H. heurippa* were necessary for the stimulation of courtship (Fig. 3). *H. heurippa* males were less than half as likely to approach and court the *H. m. melpomene* or the *H. cydno cordula* pattern than their own (Fig. 3). When either the red or yellow bands were experimentally removed from the *H. heurippa* pattern, this led to a similar reduction in its attractiveness, demonstrating that both hybrid elements are necessary for mate recognition by male *H. heurippa* (Fig. 3).

Similar results were obtained when these experiments were replicated with printed-paper models (Fig. 3), showing that the colour pattern itself was the cue rather than pheromones associated with the dissected wings. Additional experiments showed that males of both *H. m. melpomene* and *H. cydno cordula* showed a greatly reduced probability of approaching and courting the *H. heurippa* pattern than their own (Supplementary Figs 6 and 7). Given the incomplete postzygotic reproductive isolation between all three species¹, this pattern-based assortative mating must have a continuing role in generating reproductive isolation between *H. heurippa* and its relatives.

Novel patterns in *Heliconius* probably become established through a combination of genetic drift and subsequent fixation of the novel pattern driven by frequency-dependent selection²⁵. Such an event could have established the hybrid *H. heurippa* pattern as a geographic isolate of *H. cydno*. Subsequently, the pattern was sufficiently distinct from both *H. melpomene* and *H. cydno* that mate-finding behaviour also diverged in parapatry, generating assortative mating between all three species (Supplementary Fig. 8). This two-stage process indicates a possible route by which the theoretical difficulty of a rapid establishment of reproductive isolation between the hybrid and the parental taxa could have been overcome^{5,6}. Furthermore, because we are proposing divergence in mate behaviour in a geographically isolated population, reinforcement or some other form of sympatric divergence is not required for speciation to occur.

Our study provides the first experimental demonstration of a hybrid trait generating reproductive isolation between animal species, and the first example of a hybrid trait causing pre-mating isolation through assortative mating. None of the theoretical treatments of homoploid hybrid speciation have considered the effects of assortative mating^{5,6}. If variation for mate preference were incorporated, the theoretical conditions favouring hybrid speciation might

not be as stringent as has been supposed. Finally, two other species, *H. pacheus*²⁰ and *H. timareta*²⁶, have also been proposed as having *H. cydno*/*H. melpomene* hybrid patterns, indicating that this process might have occurred more than once. However, whether these cases represent a particularity of *Heliconius* or a common natural process that has been undetected in other animal groups studied less intensively remains a matter of further study. Suggestively, other proposed cases of homoploid hybrid speciation in animals occur in well-studied groups such as African cichlids^{8–10} and *Rhagoletis* flies¹¹.

METHODS

Crosses. Crosses were performed in La Vega, Colombia, between January 2000 and December 2002 with the use of *H. m. melpomene* (Virgen de Chirajara, 4.213° N, 73.795° W) and *H. cydno cordula* (Barro Negro, 6.016° N, 72.091° W), both from the Colombian Eastern Cordillera. We isolated virgin females with older males to produce inter-specific F₁ offspring and backcrosses. After mating, females were kept individually in 2 × 3 × 2 m³ insectaries and supplied with pollen and nectar from *Psiguria* and *Lantana* flowers, and *Passiflora* vines for oviposition²². Eggs and the larval and adult stages were reared as described

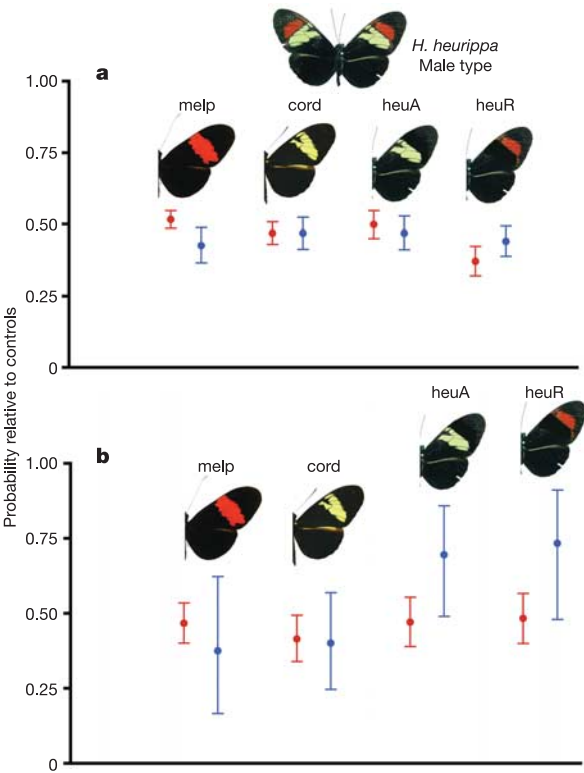


Figure 3 | Relative probabilities of *H. heurippa* males approaching and courting colour pattern models. Values are estimated relative to the probability of approach (a) and courtship (b) of the *H. heurippa* control pattern shown at the top (probability equal to 1). Model patterns are *H. m. melpomene* (melp), *H. cydno cordula* (cord), *H. heurippa* modified to remove the red band (heuA) and *H. heurippa* modified to remove the yellow band (heuR). Red, real wings; blue, paper wings. Error bars show maximum-likelihood support limits (see Supplementary Information for an explanation).

Table 2 | Number of matings in tetrad mate-choice experiments

Female	Male	
	<i>H. melpomene</i>	<i>H. heurippa</i>
<i>H. melpomene</i>	15	0
<i>H. heurippa</i>	0	12
	<i>H. cydno</i>	<i>H. heurippa</i>
<i>H. cydno</i>	5	3
<i>H. heurippa</i>	0	5
	<i>H. melpomene</i>	<i>H. cydno</i>
<i>H. melpomene</i>	10.5	0
<i>H. cydno</i>	0	5.5

Mating results of 0.5 are due to simultaneous mating of both pairs during the experiment.

previously²³. Colour pattern segregation and designation were studied with the use of nomenclature described previously²².

No-choice experiments. Experiments without choice measure the reluctance of males and females to mate inter-specifically; such experiments simulate a natural situation in which males encounter females singly. In the insectaries mentioned above, a virgin female (one to three days old) of each species was presented to ten mature males (more than ten days old) for two days. Males were used only once. Matings were recorded every 30 min from 06:00 until 15:00. Experiments were performed between all combinations of the three species, including control experiments between conspecifics. To detect any unobserved mating, all the females were checked after the experiments for the presence of a spermatophore in their reproductive tract.

Tetrads. Experiments with choice were performed to estimate mating probability in a situation in which males encounter females from different species simultaneously. Pairwise experiments were performed between each combination of the three species, in which a single recently emerged virgin female and a mature male (more than ten days old) of each species were placed in a $2 \times 3 \times 2 \text{ m}^3$ insectary over the course of a single day. Thus, each experiment involved four butterflies, for example a female and male *H. heurippa* and a female and male *H. cydno* for the comparison between those two species. The first mating only was recorded for each experiment, and individuals were not reused. In cases in which both pairs mated simultaneously they were scored as each having half a mating. At least 15 experiments were performed for each pairwise comparison. Mating probabilities were estimated by likelihood (Supplementary Methods).

Colour pattern models. In *Heliconius*, males use colour patterns to locate females and are choosy in mating, presumably because of the large and costly spermatophore transferred to females²⁷. We investigated male preferences for different colour pattern models. Between 10 and 20 males in a $2 \times 3 \times 2 \text{ m}^3$ insectary were presented with either dissected natural wings or a printed colour pattern model, fixed to a length of flexible clear nylon. Models were manipulated to simulate *Heliconius* flight in the centre of a spherical area (60 cm in diameter) demarcated by references in the insectary roof. Randomly ordered pairs of 5-min experiments were performed: first, a control flight with a model of the male's own colour pattern, and second, an experimental flight with a different colour pattern. Entry to the sphere was recorded as 'approach' and sustained fluttering directed at the model as 'courtship'. At least 25 replicates were performed for each comparison. In addition, models were made in which the *H. heurippa* pattern was modified to show either the yellow band without red (Heu-A) or the red band without yellow (Heu-R). For the dissected wing models, permanent black marker pen (Pilot ultrafine point no xylene SCA-UF) was used to cover the corresponding band. Paper models were made from digital photographs of wings taken with a Sony Cyber-shot dsc-s85 camera that were printed with a high-performance inkjet printer (Hewlett Packard Deskjet 3820) on special photo-quality calcium paper. Only paper models with a reflectance spectra similar to real wings were used. The data were used to estimate the probabilities Q_{ij} that males of type j approached or courted models of type i relative to that of their own type j (which was set to one), using likelihood. Confidence intervals for parameters were obtained as the values that decreased the difference between two log-likelihoods by two units (Supplementary Methods).

Microsatellites. Twelve microsatellite loci were genotyped (Hel02, Hel04, Hel05, Hm01, Hm02, Hm03, Hm04, Hm05, Hm06, Hm13, Hm19 and Hm22)²⁸ for 60 individuals of *Heliconius heurippa* and at least 24 individuals from each of five populations of both *H. melpomene* and *H. cydno*. Collection sites were as follows: *H. c. chioneus* and *H. m. rosina* from Pipeline Road, Panama (9.122° N, 79.715° W); *H. c. cordula*, *H. m. melpomene* from San Cristóbal, Venezuela (7.767° N, 72.225° W); *H. c. chioneus* and *H. m. melpomene* from Parcela 33, Colombia (5.066° N, 74.561° W); *H. heurippa* and *H. m. melpomene* from near Villavicencio, Colombia (4.151° N, 73.635° W); *H. c. weymeri* and *H. c. cydnides* from Ocache, Colombia (3.703° N, 76.493° W); *H. c. barinasensis* and *H. m. melpomene* from La Gira, Venezuela (9.334° N, 70.730° W). The genetic structure of populations was analysed with bayesian assignment tests as implemented in Structure 2.1 (ref. 29).

Received 8 November 2005; accepted 22 March 2006.

- Salazar, C. A. et al. Hybrid incompatibility is consistent with a hybrid origin of *Heliconius heurippa* Hewitson from its close relatives, *Heliconius cydno* Doubleday and *Heliconius melpomene* Linnaeus. *J. Evol. Biol.* **18**, 247–256 (2005).
- Rieseberg, L. H. Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* **28**, 359–389 (1997).
- Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, USA, 2004).
- Gross, B. L. & Rieseberg, L. H. The ecological genetics of homoploid hybrid speciation. *J. Hered.* **96**, 241–252 (2005).

- McCarthy, E. M., Asmussen, M. A. & Anderson, W. W. A theoretical assessment of recombinational speciation. *Heredity* **74**, 502–509 (1995).
- Buerkle, C. A., Morris, R. J., Asmussen, M. A. & Rieseberg, L. H. The likelihood of homoploid hybrid speciation. *Heredity* **84**, 441–451 (2000).
- Rieseberg, L. H. Crossing relationships among ancient and experimental sunflower hybrid lineages. *Evolution Int. J. Org. Evolution* **54**, 859–865 (2000).
- Salzburger, W., Baric, S. & Sturmbauer, C. Speciation via introgressive hybridization in East African cichlids? *Mol. Ecol.* **11**, 619–625 (2002).
- Smith, P. F., Konings, A. & Kornfield, I. Hybrid origin of a cichlid population in Lake Malawi: implications for genetic variation and species diversity. *Mol. Ecol.* **12**, 2497–2504 (2003).
- Seehausen, O. Hybridization and adaptive radiation. *Trends Ecol. Evol.* **19**, 198–207 (2004).
- Schwarz, D., Matta, B. M., Shakir-Botteri, N. L. & McPherson, B. A. Host shift to an invasive plant triggers rapid animal hybrid speciation. *Nature* **436**, 546–549 (2005).
- Brown, K. S. The biology of *Heliconius* and related genera. *Annu. Rev. Entomol.* **26**, 427–456 (1981).
- Brown, K. S., Emmel, T. C., Eliazar, P. J. & Suomalainen, E. Evolutionary patterns in chromosome numbers in neotropical Lepidoptera. I. Chromosomes of the Heliconiini (Family Nymphalidae: Subfamily Nymphalinae). *Hereditas* **117**, 109–125 (1992).
- McMillan, W. O., Jiggins, C. D. & Mallet, J. What initiates speciation in passion-vine butterflies? *Proc. Natl Acad. Sci. USA* **94**, 8628–8633 (1997).
- Mallet, J., McMillan, W. O. & Jiggins, C. D. in *Endless Forms. Species and speciation* (eds Howard, D. J. & Berlocher, S. H.) 390–403 (Oxford Univ. Press, New York, 1998).
- Jiggins, C. D., Naisbit, R. E., Coe, R. L. & Mallet, J. Reproductive isolation caused by colour pattern mimicry. *Nature* **411**, 302–305 (2001).
- Jiggins, C. D., Estrada, C. & Rodrigues, A. Mimicry and the evolution of premating isolation in *Heliconius melpomene* Linnaeus. *J. Evol. Biol.* **17**, 680–691 (2004).
- Mallet, J. & Gilbert, L. E. Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in *Heliconius* butterflies. *Biol. J. Linn. Soc.* **55**, 159–180 (1995).
- Smiley, J. T. Plant chemistry and the evolution of host specificity: new evidence from *Heliconius* and *Passiflora*. *Science* **201**, 745–747 (1978).
- Gilbert, L. E. in *Butterflies: Ecology and Evolution Taking Flight* (eds Boggs, C. L., Watt, W. B. & Ehrlich, P. R.) 281–318 (Univ. of Chicago Press, Chicago, 2003).
- Linares, M. *Adaptive Microevolution Through Hybridization and Biotic Destruction in the Neotropics*. Thesis, University of Texas, Austin (1989).
- Naisbit, R. E., Jiggins, C. D. & Mallet, J. Mimicry: developmental genes that contribute to speciation. *Evol. Dev.* **5**, 269–280 (2003).
- Naisbit, R. E., Jiggins, C. D., Linares, M., Salazar, C. A. & Mallet, J. Hybrid sterility, Haldane's rule and speciation in *Heliconius cydno* and *H. melpomene*. *Genetics* **161**, 1517–1526 (2002).
- Mallet, J. & Barton, N. H. Strong natural selection in a warning-color hybrid zone. *Evolution Int. J. Org. Evolution* **43**, 421–431 (1989).
- Mallet, J. & Singer, M. C. Individual selection, kin selection and the shifting balance in the evolution of warning colours: the evidence from butterflies. *Biol. J. Linn. Soc.* **32**, 337–350 (1987).
- Mallet, J. Causes and consequences of a lack of coevolution in Müllerian mimicry. *Evol. Ecol.* **13**, 777–806 (1999).
- Boggs, C. L. & Gilbert, L. E. Male contribution to egg production in butterflies: evidence for transfer of nutrients at mating. *Science* **206**, 83–84 (1979).
- Mavárez, J. & González, M. A set of microsatellite loci for *Heliconius melpomene* and close relatives. *Mol. Ecol. Notes* **6**, 20–23 (2006).
- Pritchard, J. K., Stephens, M. & Donnelly, P. J. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. García and the UNET for help at Paramillo Natural Park, San Cristóbal, Venezuela; R. Castillo, L. Pereira and O. Quintero for butterfly collecting; M. Guerra and L. González for help with the preparation of figures; N. Barton and F. Jiggins for discussion; and L. Gilbert and J. Mallet for inspiring us to study hybridization. This work was funded by the Marie-Curie Fellowships, the Smithsonian Tropical Research Institute, the Fondo Colombiano de Investigaciones Científicas y Proyectos Especiales Francisco José de Caldas COLCIENCIAS, Banco de la República, and private donations from Continuos S.A., Profol El Carmen S.A., Didacol S.A., and F. Arango, Colombia. C.D.J. is supported financially by the Royal Society and by a grant from BBSRC.

Author Information The sequences have been deposited in GenBank under accession numbers DQ445384–DQ445414 (*Distal-less*) and DQ445416–DQ445457 (*Inverted*). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.L. (mlinares@uniandes.edu.co) or J.M. (mavarezj@si.edu).

LETTERS

Sperm storage induces an immunity cost in ants

Boris Baer^{1*†}, Sophie A. O. Armitage^{1*} & Jacobus J. Boomsma¹

Ant queens are among the most long-lived insects known^{1,2}. They mate early in adult life and maintain millions of viable sperm in their sperm storage organ until they die many years later^{3,4}. Because they never re-mate, the reproductive success of queens is ultimately sperm-limited, but it is not known what selective forces determine the upper limit to sperm storage. Here we show that sperm storage carries a significant cost of reduced immunity during colony founding. Newly mated queens of the leaf-cutting ant *Atta colombica* upregulate their immune response shortly after completing their nest burrow, probably as an adaptive response to a greater exposure to pathogens in the absence of grooming workers. However, the immune response nine days after colony founding is negatively correlated with the amount of sperm in the sperm storage organ, indicating that short-term survival is traded off against long-term reproductive success. The immune response was lower when more males contributed to the stored sperm, indicating that there might be an additional cost of mating or storing genetically different ejaculates.

Delayed internal fertilization has evolved multiple times⁵ and allows sperm competition and cryptic female choice between ejaculates^{6–8}. Sperm storage imposes selection on males to increase sperm number and viability, directly or indirectly by means of the seminal fluids, and selection on females to maintain ejaculates⁹. However, the selective forces that have shaped sperm storage traits in ants are not primarily related to sexual selection, as in most other animals, but to iteroparous perennial reproduction and extreme reproductive division of labour¹⁰. Ant queens obtain their lifetime supply of sperm on a single mating flight. When they succeed in establishing a colony they may live for several decades, whereas males die on the day of mating but ‘survive’ as stored sperm. Queens never re-mate later in life, which implies that they are not promiscuous in the usual sense even when mating with multiple males. As the lifetime reproductive success of ant queens is ultimately constrained by the availability of stored sperm, there has been strong selection for keeping sperm viable and using only a few sperm to fertilize each egg¹¹. However, it has remained unclear what factors have constrained the evolution of larger sperm stores. Although mating with more males might incur prohibitive costs of longer exposure to predators and diseases, it seems hard to understand why ant males have not evolved larger ejaculates that would allow queens to store more sperm without having to increase the number of matings¹².

We propose that the long-term storage and maintenance of large quantities of sperm incurs a significant metabolic cost and that this cost is a major factor explaining sperm limitation in ant queens. Suggestive indications of such costs have been reported for the honeybee (*Apis mellifera*)³ and the ant *Crematogaster opuntiae*³, but this idea has never been tested explicitly. A complication is that the cost of sperm storage may not significantly affect the fitness of queens in established colonies. During colony founding, however, solitary queens are severely resource-limited until their first workers emerge, so that metabolic costs of maintaining large quantities of stored

sperm may interfere with other vital functions. This implies that evolutionary trade-offs may become apparent that are not expressed later in life when workers provide ample resources and care.

In contrast to established queens, founding queens are exposed to a variety of soil-borne pathogens that would be unlikely to reach them in the protected nest environment of mature colonies¹³. Mortality rates are typically more than 95% during colony founding^{14,15}, with pathogens accounting for 74% of failures during colony foundation in the leaf-cutting ant *Atta cephalotes*¹⁵. The importance of a well-functioning immune system early in colony life has been documented for bumblebees (*Bombus terrestris*)^{16,17}, in which the encapsulation response of the first worker brood is a significant predictor for eventual colony size and fitness. There is also evidence that maintaining immune responses in insects is costly¹⁸. We therefore expect that founding ant queens should increase their investment into immune defence shortly after mating to avoid premature death from fungal or bacterial infections, and that this adaptive response might be compromised by the metabolic costs of sperm storage in this crucial solitary period of a queen's life.

We tested this hypothesis on founding queens of the Panamanian leaf-cutting ant *A. colombica*, which are among the most long-lived and fertile insects known. Queens store up to several hundred million sperm, usually from two or three males¹⁹. After dispersal and mating they dig a burrow and seal it. Soon afterwards they lay their first eggs and start building a fungus garden from a fragment of mycelium that they take along from their maternal colony¹⁴ (Fig. 1a). In the months to follow, they nurse their brood and incipient fungus garden with trophic eggs and faecal droplets, while recycling their flight-muscle proteins and abdominal fat reserves until the first workers eclose and start foraging²⁰. Once founding colonies have reached this stage, they may grow to contain several million workers and will produce yearly clutches of thousands of winged new queens and males until they die one or several decades later¹⁴.

As expected, immune defence measured as encapsulation response (see Methods and Supplementary Information) was significantly higher ($t = -3.299$, d.f. = 44, $P = 0.002$) in queens that were excavated nine days after the mating flight (48.61 ± 4.48) than in queens excavated after one day (30.69 ± 3.08 ; Fig. 1b). After nine days, the mortality of queens in their burrows had increased to 21%. This corresponds to an average of 2.6% per day, very close to the 3% mortality observed in the queen sample taken after one day (Fig. 1b) and consistent with 95% mortality being reached after about 3.5 months, in agreement with previous estimates^{14,15}. Additional data on founding queens from the same site in the following year showed that the average encapsulation response and number of haemocytes (another measure of the potential of immune defence) one day after mating were not significantly different from values obtained for winged virgin queens before mating (encapsulation response: $F_{1,67} = 1.771$, $P = 0.188$; haemocyte number: $F_{1,67} = 1.536$, $P = 0.220$), whereas both responses were significantly elevated nine days after mating (multivariate analysis of covariance $F_{2,62} = 6.618$,

¹Institute of Biology, Department of Population Biology, Universitetsparken 15, 2100 Copenhagen, Denmark. [†]Present address: Zoology Building, School of Animal Biology (MO92), The University of Western Australia, 6009 Nedlands, Western Australia, Australia.

*These authors contributed equally to this work.

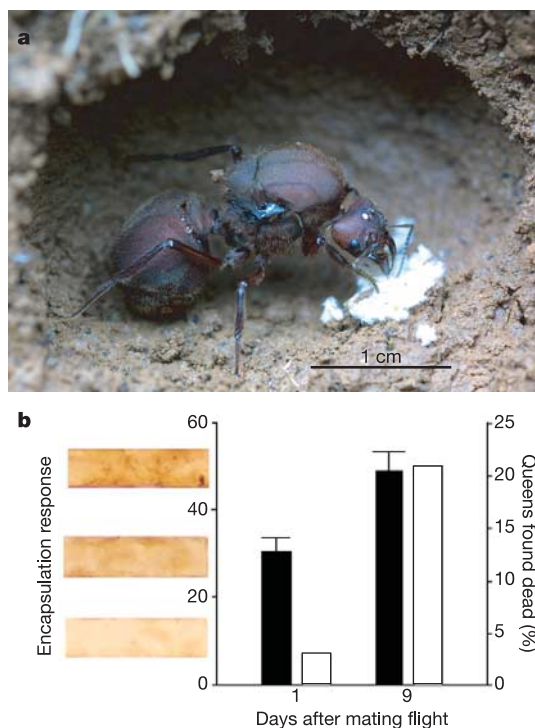


Figure 1 | Increase in immune response during colony founding. **a**, A large (about 2.2 cm) founding queen of *A. colombica* in her initial brood chamber with the incipient fungus garden. **b**, The encapsulation response (filled bars; means $\pm$ s.e.m. for 23 queens each) and mortality rates (open bars) of colony-founding queens one and nine days after their mating flight. Horizontal bars at the left are representative examples of melanization of the nylon inserts that were used to quantify encapsulation (see Methods and Supplementary Information). Encapsulation differed significantly between queens collected one and nine days after their mating flight (see the text), and this difference remained significant when including queen weight (not significant) as a covariate in the analysis ($F_{1,43} = 11.712$, $P = 0.001$).

$P = 0.002$; see Supplementary Information). This indicates that the encapsulation response is indeed upregulated after colony founding and not downregulated during the mating flight. These results are consistent with a recent gene expression study in *Solenopsis* fire ants, in which two precursors of antibacterial peptides were shown to be upregulated in queens one day after their mating flight²¹.

It was not possible to count stored sperm in dead queens, but the 46 queens sampled alive had stored up to 465×10^6 sperm. The overall mean was $(244 \pm 12) \times 10^6$ sperm (mean $\pm$ s.e.m.) per queen, with no significant difference for queens sampled one and nine days after mating ($t = 0.025$, d.f. = 44, $P = 0.981$). For the queens collected after nine days, 65% of the variation in encapsulation response could be explained by the amount of stored sperm, whereas no such relationship was present in the sample of founding queens taken one day after mating (Fig. 2). Queens with higher than average sperm stores ($(313 \pm 14) \times 10^6$) were apparently unable to increase their immune response (difference between day 1 and day 9 after founding for 22 queens with more than 244×10^6 sperm: $t = -0.637$, $P = 0.532$), whereas the remaining 24 queens with sperm stores lower than average ($(181 \pm 7) \times 10^6$) increased their encapsulation response by 100% (from 29.86 to 59.49 on average) between day 1 and day 9 ($t = -4.922$, $P < 0.0001$).

Microsatellite data for the number of father haplotypes represented in the stored sperm showed that queens had mated with two to five males (see Supplementary Information for details), confirming earlier data obtained from the same population¹⁹. The expected positive correlation between the inferred number of fathers and the number of stored sperm was weak ($r = 0.211$; one-tailed $P = 0.085$;

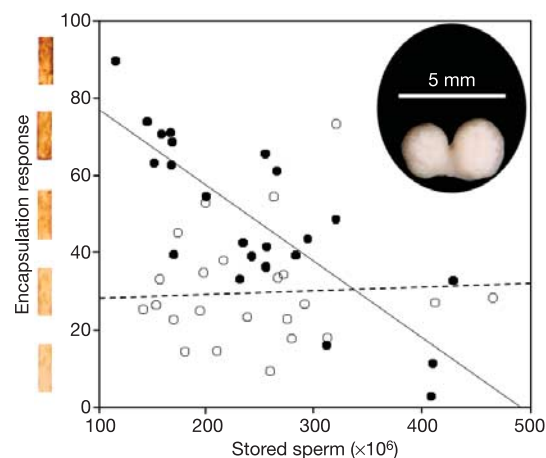


Figure 2 | Effect of stored sperm on immune response. The encapsulation response (indicated with bars along the y axis as in Fig. 1) as a function of the amount of sperm stored in a queen's spermatheca. Open circles refer to queens collected one day after mating, at which time there is no relationship between the two variables (dashed linear regression line: $F_{1,21} = 0.063$, $P = 0.804$, $r^2 = 0.003$). Filled circles refer to the encapsulation responses of queens collected nine days after mating. At this time mean encapsulation responses were elevated by 63% (from 30.7 to 48.6 on average; Fig. 1) and showed a significant negative correlation with the number of sperm stored in the spermatheca (solid regression line: $F_{1,21} = 39.282$, $P < 0.0001$, $r^2 = 0.652$). The inset illustrates the large size of the sperm storage organ (white) relative to the approximate outline of a queen gaster (black).

$n = 44$), which is consistent with a previous study¹⁹. However, the correlation between immune response and the inferred number of fathers was negative and significant ($r = -0.314$; $P = 0.038$; $n = 44$).

In a multiple regression analysis with backwards elimination of non-significant terms, all but one of the interaction terms and queen weight ($F_{1,38} = 0.083$, $P = 0.775$) were excluded. This produced an overall model for encapsulation response with a significant positive effect of day of sampling (day 1 versus day 9; $F_{1,39} = 31.334$, $P < 0.001$, $r^2 = 0.45$), significant negative effects of both the number of stored sperm ($F_{1,39} = 10.410$, $P = 0.003$, $r^2 = 0.21$) and the number of fathers ($F_{1,39} = 5.854$, $P = 0.020$, $r^2 = 0.13$), and a significant sampling-day $\times$ sperm-number interaction term ($F_{1,39} = 18.886$, $P < 0.001$, $r^2 = 0.33$), showing that all singular relationships (Figs 1 and 2) were robust in partial analysis and that being inseminated by more males induced a significant additional immunity cost (Fig. 3).

Our results indicate that a founding queen's risk of dying from disease might be proportional to the amount of sperm stored during her mating flight because maintaining stored sperm carries considerable metabolic costs, which trade off against the needs of immune defences during solitary colony founding. A corollary would be that the average number of sperm stored by a queen that has survived the colony-founding stage is lower than the amount observed in an average newly mated queen, because of differential mortality related to the cost of sperm storage. This seems to be true, because a sample from the same site of young *A. colombica* queens that had already produced their first hundreds of workers²² had significantly smaller numbers of stored sperm ($(124 \pm 11) \times 10^6$; range $(57-255) \times 10^6$) than the average of 244×10^6 observed in the present study. Storing more sperm might thus increase the risk of failure early in life but be beneficial later, when the prolonged ability to fertilize eggs might add several years of colony life and reproduction. However, our results are not consistent with the assumption that the sperm affects the immune system directly, because this should have implied an increase in immune response with the number of stored sperm.

Atta leaf-cutting ants are among the few clades of social insects that

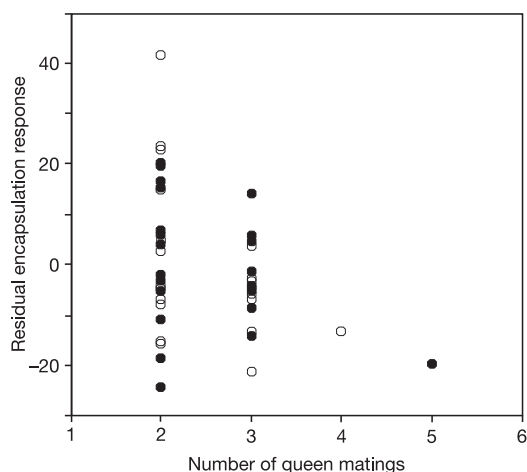


Figure 3 | Effect of number of mates on immune response. Residual encapsulation responses based on the regressions from day 1 (open circles) and day 9 (filled circles) in Fig. 2, plotted as a function of the number of fathers that contributed to the stored sperm. Differences between day 1 and day 9 for queens that mated with two males ($t = 0.662$, $P > 0.5$) or with three to five males ($t = 1.047$, $P > 0.2$) were not significant (tested separately because of unequal variances). The overall partial regression ($F_{1,38} = 5.396$, $P = 0.026$, $r^2 = 0.12$) across the pooled data set was significant (see the text). When distinguishing between only two fathers and more than two fathers, the average residual encapsulation response for queens that mated with two males ($+3.403$; $n = 25$) and queens that mated with three to five males (-4.680 ; $n = 19$) remained significantly different (after adjustment for unequal variances: $t = 2.144$, $P = 0.040$).

have obligatory multiple mating of queens^{10,23}. Because colonies need several years of ergonomic growth to have enough workers to start reproducing, it has been proposed that sexually transmitted diseases should be absent or non-virulent because they would otherwise eliminate their own transmission¹³. Similarly, male accessory gland compounds that would reduce the survival of competing ejaculates or have directly negative effects on female fitness later in life are expected to be of limited significance in social insects^{10,23,24}. Our finding that storing a specific amount of sperm of n males is more costly than storing the same amount of sperm from $n - 1$ males (Fig. 3) is therefore interesting in two ways. It may indicate that male–male conflicts over sperm storage have not been completely eliminated, in spite of queen mates being lifetime committed and having a joint interest in building a strong colony of sterile workers before reproduction can start some years later. These conflicts may therefore be similar to those social insects for which single mating of the queen is the rule, as recent work on bumblebees has shown that mixed artificially inseminated sperm from more than one male significantly reduced queen fitness relative to the same number of sperm from a single male²⁵. Alternatively, the negative effect of the number of inseminations on immune response in founding queens may reflect a delayed cost of mating, because mating with more males is likely to imply more hours of flight activity so that fewer resources may be left to mount immune defences during subsequent colony founding. This possibility is intriguing because this type of mating cost has never been empirically quantified in any social insect.

METHODS

Sampling. On 18 May 2004 we marked 120 entrance holes of *A. colombica* burrows on a single lawn in Gamboa, Panama, about 30 h after a nocturnal mating flight. A random subset of 38 of these queens was excavated immediately afterwards and taken to the laboratory, where each queen was maintained at ambient temperature in a Petri dish with moist cotton wool. For a random sample of 23 of these queens, the encapsulation response over a 24-h period was measured, and sperm counts and microsatellite paternity data of the stored sperm were obtained (see below). Nine days after the mating flight a second sample of 53 live queens was excavated, of which we used a random subset of 23

for the same measurements. Mortality after one and nine days was assessed at excavation (see also Supplementary Information).

Measuring the immune response. We estimated the efficiency of immune defence by experimentally eliciting an encapsulation response (see Supplementary Information for a more extensive overview of this technique). This was done by inserting a small piece of nylon (0.13 mm $\times$ 0.50 mm) through the queen's intersegmental membrane between the 6th and 7th sternites. For the implantation we adapted machinery normally used for artificially inseminating bumblebees²⁵. After 24 h the queens were freeze-killed and the encapsulated nylon filaments were dissected out and mounted on microscope slides, then photographed with a Canon EOS D30 digital camera connected to a Leica stereomicroscope at $\times 62.5$ magnification. Shades of grey measure the deposition of dead melanized haemocytes. Melanin is synthesized from the amino acid tyrosine, a reaction that is in part catalysed by the enzyme phenoloxidase, producing reactive intermediates of oxygen and nitrogen that have been implicated in parasite death (see Supplementary Information for further details). The extent of grey coloration on the implants was analysed with the Image 1.62 software developed by the US National Institutes of Health (<http://rsb.info.nih.gov/nih-image/>). To verify the generality of our 2004 estimations, encapsulation responses after one and nine days were also obtained from founding queens in 2005, and compared with encapsulation responses of virgin queens just before mating, and also with haemocyte counts for the same individuals to examine whether alternative measures of immune response gave similar results (see Supplementary Information).

Sperm counts. Spermathecae were dissected out of frozen queens and transferred to 2 ml of culture medium for boar semen (Merk III diluant, purchased from Minitüb). They were ruptured with watchmaker forceps after which solvent and sperm were mixed by vortex-mixing the solution for 20–30 s. Subsamples (50 μ l) of this stock solution were transferred to 1.95 ml of distilled water and vortex-mixed again for 20–30 s. Single drops of 1 μ l of this solution were placed on a microscope slide and air-dried. Sperm counts were performed after staining with 4',6-diamidino-2-phenylindole (Merck) by counting all fluorescence-marked sperm heads within a single 1- μ l drop of diluted sperm, using an Olympus fluorescence microscope at $\times 400$ magnification.

Number of matings and statistics. The number of fathers that contributed to the stored sperm was estimated after amplifying two highly variable microsatellite loci Etta5-6TF and Etta7-8TF (ref. 26) from both a sample of the stock solution of sperm and a leg of each corresponding queen that the sperm came from. Reliable estimates of the number of inseminations were obtained for 44 of the 46 queens (see Supplementary Information for details), so mate number could be used as an additional predictor for encapsulation response in multiple regression analysis with stored sperm, queen weight and date of excavation as other predictor variables. The listed r^2 values for the multiple regression analysis refer to partial proportions of explained variance. All means are given with s.e.m.

Received 24 January; accepted 7 March 2006.

- Keller, L. & Genoud, M. Extraordinary lifespans in ants: a test of evolutionary theories of ageing. *Nature* **389**, 958–960 (1997).
- Pamilo, P. Life span of queens in the ant *Formica exsecta*. *Insect. Soc.* **38**, 111–120 (1991).
- Wheeler, D. E. & Krutzsch, P. H. Ultrastructure of the spermatheca and its associated gland in the ant *Crematogaster opuntiae* (Hymenoptera, Formicidae). *Zoomorphology* **114**, 203–212 (1994).
- Hölldobler, B. & Bartz, S. H. in *Experimental Behavioural Ecology and Sociobiology* (eds Hölldobler, B. & Lindauer, M.) 237–257 (Gustav Fischer, Stuttgart, 1985).
- Birkhead, T. R. & Möller, A. P. *Sperm Competition and Sexual Selection* (Academic, New York, 1998).
- Eberhard, W. G. *Female Control: Sexual Selection by Cryptic Female Choice* (Princeton Univ. Press, Princeton, 1996).
- Simmons, L. W. *Sperm Competition and its Evolutionary Consequences in the Insects* (Princeton Univ. Press, Princeton, 2001).
- Baer, B. Sexual selection in *Apis* bees. *Apidologie (Celle)* **36**, 187–200 (2005).
- Breque, C., Surai, P. & Brillard, J. P. Roles of antioxidants on prolonged storage of avian spermatozoa *in vivo* and *in vitro*. *Mol. Reprod. Dev.* **66**, 314–323 (2003).
- Boomsma, J. J., Baer, B. & Heinze, J. The evolution of male traits in social insects. *Annu. Rev. Entomol.* **50**, 395–420 (2005).
- Tschinkel, W. R. & Porter, S. D. Efficiency of sperm use in queens of the fire ant, *Solenopsis invicta* (Hymenoptera: Formicidae). *Ann. Entomol. Soc. Am.* **81**, 777–781 (1988).
- Crozier, R. H. & Page, R. E. On being the right size: Male contributions and multiple mating in social hymenoptera. *Behav. Ecol. Sociobiol.* **18**, 105–116 (1985).
- Boomsma, J. J., Schmid-Hempel, P. & Hughes, W. O. H. in *Insect Evolutionary Ecology* (eds Fellowes, M., Holloway, G. & Roloff, J.) 139–175 (CABI, Wallingford, 2005).

14. Weber, N. A. *Gardening Ants: The Attines* (American Philosophical Society, Philadelphia, 1972).
15. Fowler, H. G., Pereira da Silva, V., Forti, L. C. & Saes, N. B. in *Fire Ants and Leaf Cutting Ants: Biology and Management* (eds Lofgren, C. S. & Vander Meer, R. K.) 123–145 (Westview Press, Boulder, Colorado, 1986).
16. Baer, B. & Schmid-Hempel, P. Effects of selective episodes in the field on life history traits in the bumblebee *Bombus terrestris*. *Oikos* **101**, 563–568 (2003).
17. Baer, B. & Schmid-Hempel, P. Sperm in fluences female hibernation success, survival and fitness in the bumblebee *Bombus terrestris*. *Proc. R. Soc. B* **272**, 319–323 (2005).
18. Rolff, J. & Siva-Jothy, M. T. Invertebrate ecological immunology. *Science* **301**, 472–475 (2003).
19. Fjerdingstad, E. J. & Boomsma, J. J. Multiple mating increases the sperm stores of *Atta colombica* leafcutter ant queens. *Behav. Ecol. Sociobiol.* **42**, 257–261 (1998).
20. Hölldobler, B. & Wilson, E. O. *The Ants* (Springer, Berlin, 1990).
21. Tian, H. S., Vinson, S. B. & Coates, C. J. Differential gene expression between alate and dealate queens in the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae). *Insect Biochem. Mol. Biol.* **34**, 937–949 (2004).
22. Fjerdingstad, E. J. & Boomsma, J. J. Variation in size and sperm content of sexuals in the leafcutter ant *Atta colombica*. *Insect. Soc.* **44**, 209–218 (1997).
23. Baer, B. & Boomsma, J. J. Male reproductive investment and queen mating frequency in fungus growing ants. *Behav. Ecol.* **15**, 426–432 (2004).
24. Baer, B. & Schmid-Hempel, P. Unexpected consequences of polyandry for parasitism and fitness in the bumblebee, *Bombus terrestris*. *Evolution Int. J. Org. Evolution* **55**, 1639–1643 (2001).
25. Baer, B. & Schmid Hempel, P. The artificial insemination of bumblebee queens. *Insect. Soc.* **47**, 183–187 (2000).
26. Fjerdingstad, E. J., Boomsma, J. J. & Thoren, P. Multiple paternity in the leafcutter ant *Atta colombica*—a microsatellite DNA study. *Heredity* **80**, 118–126 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. R. Nash for advice on statistical issues; C. Andersen and S. Mathiasen for assistance with DNA work; S. Mathiasen and D. J. C. Kronauer for advice on DNA work; S. P. A. den Boer, M. Poulsen, A. M. Schmidt, J. Thomas and L. V. Ugelvig for assistance in the field, and P. D'Ettorre, D. R. Nash, J. S. Pedersen, B. Baer-Imhoof and L. W. Simmons for comments on the manuscript. This study was supported by the Swiss National Science Foundation (stipend for advanced scientists to B.B.), the EU Research-Training Network *INSECTS*, the Carlsberg Foundation and a Marie Curie Intra-European Fellowship (postdoctoral stipends to S.A.O.A.), the Danish Natural Science Research Council (B.B. and J.J.B.) and the Danish National Research Foundation (J.J.B. and S.A.O.A.). We thank the Smithsonian Tropical Research Institute in Panama for facilities and logistical support, and the Autoridad Nacional de Ambiente (ANAM) of Panama for issuing collection and export permits.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.B. (bcbaer@bi.ku.dk).

LETTERS

Cortical substrates for exploratory decisions in humans

Nathaniel D. Daw^{1*}, John P. O'Doherty^{2*†}, Peter Dayan¹, Ben Seymour² & Raymond J. Dolan²

Decision making in an uncertain environment poses a conflict between the opposing demands of gathering and exploiting information. In a classic illustration of this 'exploration–exploitation' dilemma¹, a gambler choosing between multiple slot machines balances the desire to select what seems, on the basis of accumulated experience, the richest option, against the desire to choose a less familiar option that might turn out more advantageous (and thereby provide information for improving future decisions). Far from representing idle curiosity, such exploration is often critical for organisms to discover how best to harvest resources such as food and water. In appetitive choice, substantial experimental evidence, underpinned by computational reinforcement learning² (RL) theory, indicates that a dopaminergic^{3,4}, striatal^{5–9} and medial prefrontal network mediates learning to exploit. In contrast, although exploration has been well studied from both theoretical¹ and ethological¹⁰ perspectives, its neural substrates are much less clear. Here we show, in a gambling task, that human subjects' choices can be characterized by a computationally well-regarded strategy for addressing the explore/exploit dilemma. Furthermore, using this characterization to classify decisions as exploratory or exploitative, we employ functional magnetic resonance imaging to show that the frontopolar cortex and intraparietal sulcus are preferentially active during exploratory decisions. In contrast, regions of striatum and ventromedial prefrontal cortex exhibit activity characteristic of an involvement in value-based exploitative decision making. The results suggest a model of action selection under uncertainty that involves switching between exploratory and exploitative behavioural modes, and provide a computationally precise characterization of the contribution of key decision-related brain systems to each of these functions.

Exploration is a computationally refined capacity, demanding careful regulation. Two possibilities for this regulation arise. On the one hand, we might expect the involvement of cognitive, prefrontal control systems¹¹ that can supervene¹² over simpler dopaminergic/striatal habitual mechanisms. On the other hand, theoretical work on optimal exploration^{1,13} indicates a more unified architecture, according to which actions can be assessed with the use of a metric that integrates both primary reward and the informational value of exploration, even in simple, habitual decision systems.

We studied patterns of behaviour and brain activity in 14 healthy subjects while they performed a 'four-armed bandit' task involving repeated choices between four slot machines (Fig. 1; see Supplementary Methods). The slots paid off points (to be exchanged for money) noisily around four different means. Unlike standard slots, the mean payoffs changed randomly and independently from trial to trial, with subjects finding information about the current worth of a slot only

through sampling it actively. This feature of the experimental design, together with a model-based analysis, allowed us to study exploratory and exploitative decisions under uniform conditions, in the context of a single task.

We asked subjects in post-task interviews to describe their choice strategies. The majority (11 of 14) reported occasionally trying the different slots to work out which currently had the highest payoffs (exploring) while at other times choosing the slot they thought had the highest payoffs (exploiting). To investigate this behaviour quantitatively, we considered RL (ref. 2) strategies for exploration. These strategies come in three flavours, differing in how exploratory actions are directed. The simplest method, known as ' ϵ -greedy', is undirected: it chooses the 'greedy' option (the one believed to be best) most of the time, but occasionally (with probability ϵ) substitutes a random action. A more sophisticated approach is to guide exploration by expected value, as in the 'softmax' rule. With softmax, the decision to explore and the choice of which suboptimal action to take are determined probabilistically on the basis of the actions' relative expected values. Last, exploration can additionally be directed by awarding bonuses in this latter decision towards actions whose consequences are uncertain: specifically, to those for which exploration will be most informative. The optimal strategy for a restricted class of simple bandit tasks has this characteristic¹, as do standard heuristics¹⁴ for exploration in more complicated RL tasks such as ours, for which the optimal solution is computationally intractable.

We compared the fit of three distinct RL models, embodying the aforementioned strategies, to our subjects' behavioural choices. All the models learned the values of actions with the use of a Kalman filter (see Supplementary Methods), an error-driven prediction algorithm that generalizes the temporal-difference learning algorithm (used in most RL theories of dopamine) by also tracking uncertainty about the value of each action. The models differed only in their choice rules. We compared models by using the likelihood of the subjects' choices given their experience, optimized over free parameters. This comparison (Supplementary Tables 1 and 2) revealed strong evidence for value-sensitive (softmax) over undirected (ϵ -greedy) exploration. There was no evidence to justify the introduction of an extra parameter that allowed exploration to be directed towards uncertainty (softmax with an uncertainty bonus): at optimal fit, the bonus was negligible, making the model equivalent to the simpler softmax. We conducted additional model fits (see Supplementary Information) to verify that these findings were not an artefact of our assumptions about the yoking of free parameters between subjects.

Having characterized subjects' behaviour computationally, we used the best-fitting softmax model to generate regressors containing value predictions, prediction errors and choice probabilities for each subject on each trial. We used statistical parametric mapping to

¹Gatsby Computational Neuroscience Unit, University College London (UCL), Alexandra House, 17 Queen Square, London WC1N 3AR, UK. ²Wellcome Department of Imaging Neuroscience, UCL, 12 Queen Square, London WC1N 3BG, UK. [†]Present address: Division of Humanities and Social Sciences, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.

*These authors contributed equally to this work.

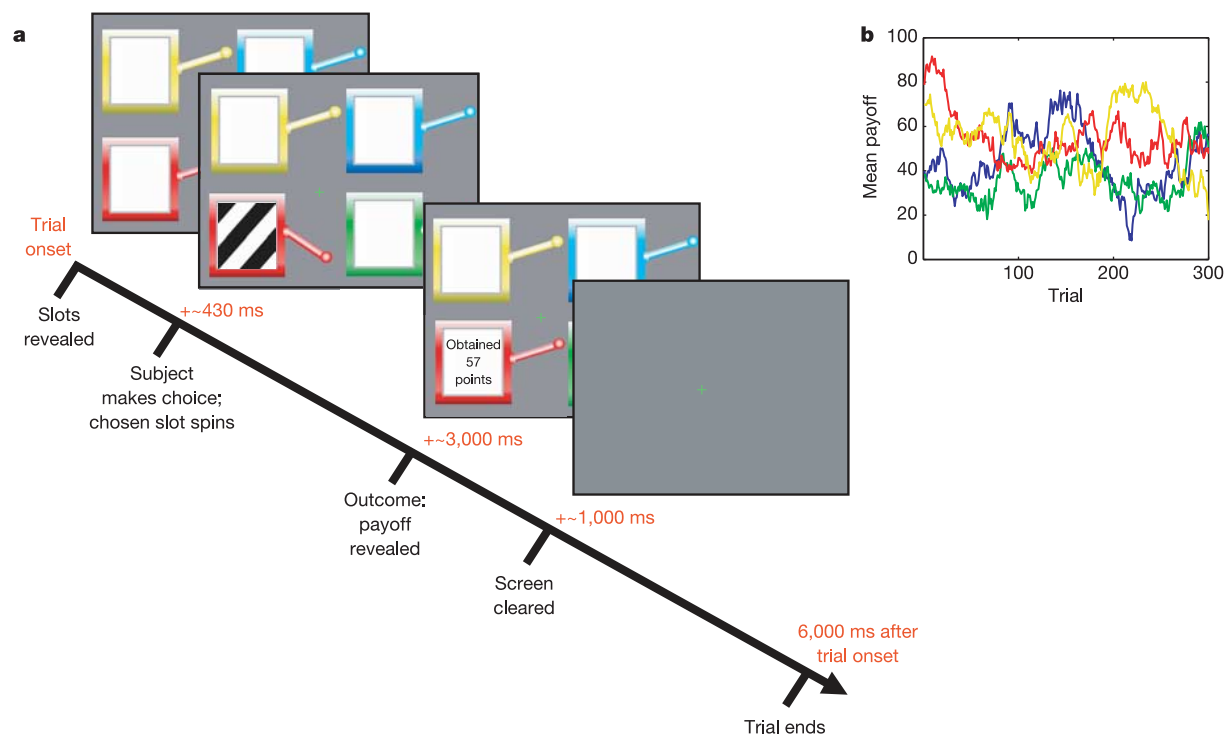


Figure 1 | Task design. **a**, Illustration of the timeline within a trial. Initially, four slots are presented. The subject chooses one, which then spins. Three seconds later the number of points won is revealed. After a further second the screen is cleared. The next trial is triggered after a fixed trial length of 6 s and an additional variable inter-trial interval (mean 2 s).

b, Example of mean payoffs that would be received for choosing each slot machine (four coloured lines) on each trial, demonstrating their independent random diffusion. The payoff received for a particular choice is corrupted by gaussian noise around this mean.

identify brain regions in which neural activity was significantly correlated with the model's internal signals. Consistent with previous studies⁷⁻⁹ was our observation that a prediction error was correlated significantly with activity in both the ventral and dorsal striatum (see

Supplementary Table 3). Other, cortical, structures linked to this subcortical network¹⁵ also showed significant value-related correlations. Specifically, we found activity in medial orbitofrontal cortex to be correlated with the magnitude of the obtained payoff (Fig. 2a), a

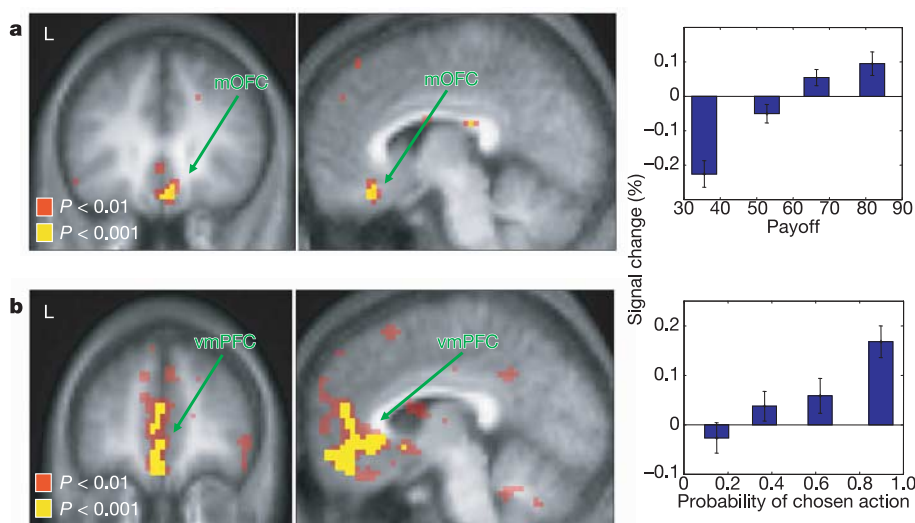


Figure 2 | Reward-related activations. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$ to illustrate the full extent of the activations) are superimposed on a subject-averaged structural scan. **a**, Region of medial orbitofrontal cortex (mOFC) correlating significantly with the number of points received. The coordinates of the activated area are [3,30,-21, peak $z = 3.87$]. The bar plot shows the average BOLD response to outcome, binned by amount won (error bars represent s.e.m.). **b**, Regions of ventromedial prefrontal cortex (vmPFC; including medial and lateral

orbitofrontal cortex and adjacent medial prefrontal cortex) correlating significantly with the probability assigned by the computational model to the subject's choice of slot. The coordinates of the activated areas are as follows: medial orbitofrontal, [-3,45,-18, peak $z = 5.62$]; lateral orbitofrontal (not illustrated), [45,36,-15, peak $z = 4.6$]; medial prefrontal, [-3,33,-6, peak $z = 4.62$]. The bar plot shows the average medial prefrontal BOLD response to decision, binned by choice probability (error bars represent s.e.m.).

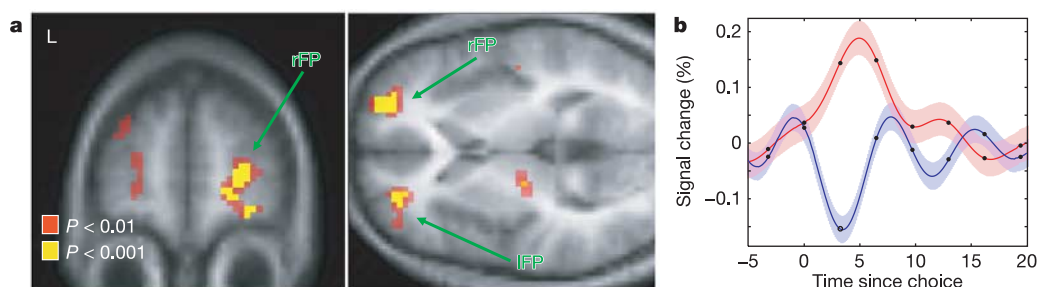


Figure 3 | Exploration-related activity in frontopolar cortex. **a**, Regions of left and right frontopolar cortex (lFP, rFP) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a subject-averaged structural scan. The coordinates of activated areas are $[-27, 48, 4, \text{peak}$

$z = 3.49]$ for lFP and $[27, 57, 6, \text{peak } z = 4.13]$ for rFP. **b**, rFP BOLD time courses averaged over 1,515 exploratory (red line) and 2,646 exploitative (blue line) decisions. Black dots indicate the sampling frequency (although, because sample alignment varied from trial to trial, time courses were upsampled). Coloured fringes show error bars (representing s.e.m.).

finding consistent with previous evidence indicating that this region is involved in coding the relative value of different reward stimuli, including abstract rewards^{16,17}. Furthermore, activity in medial and lateral orbitofrontal cortex, extending into ventro-medial prefrontal cortex, was correlated with the probability assigned by the model to the action actually chosen on a given trial (Fig. 2b). In the softmax model, this probability is a relative measure of the expected reward value of the chosen action, and the observed profile of activity is thus consistent with a role for orbital and adjacent medial prefrontal cortex in encoding predictions of future reward^{18,19}. The same quantity was negatively correlated with activity in a small area of dorsolateral prefrontal cortex (left: $-39, 36, 42$, peak $z = 3.38$; right: $36, 33, 33$, peak $z = 3.27$); that is, higher activity was seen there for lower-probability choices.

We next sought to identify brain activity that selectively reflected whether actions were chosen for their exploratory or exploitative potential. To test for such a signature, we classified trials according to whether the actual choice was the one predicted by the model to be the dominant slot machine with the highest expected value (exploitative) or a dominated machine with a lower expected value (exploratory). We then directly compared the pattern of brain activity associated with these exploratory and exploitative trials. We found no area that exhibited significantly higher activity for exploitative than exploratory decisions (employing whole-brain correction for multiple comparisons). However, the opposite contrast revealed several activations. First, right anterior frontopolar cortex (Fig. 3a) was significantly more active during decisions classified as exploratory ($P < 0.05$, corrected whole-brain for multiple comparisons with false discovery rate; activation was noted bilaterally at $P < 0.001$ uncorrected but did not survive whole-brain correction on the left). Average blood-oxygenation-level-dependent

(BOLD) signal time courses from the region (Fig. 3b) demonstrated phasic increases and decreases in activity that were time-locked to subjects' exploratory and exploitative decisions, respectively.

Because the prefrontal cortex is the principal cortical region implicated in behavioural control²⁰, the signal we observed in anterior frontopolar cortex could reflect a control mechanism facilitating the switching of behavioural strategies between exploratory and exploitative modes. This most rostral of prefrontal regions is known to be associated with high-level control²¹. This region sits atop a proposed hierarchy of nested prefrontal controllers²² and is implicated in mediating between different goals, subgoals²³ or cognitive processes²¹.

Differential activation during exploratory trials was also observed bilaterally in anterior intraparietal sulcus (whole-brain corrected at $P < 0.05$; Fig. 4), bordering on the postcentral gyrus. The sulcus has repeatedly been implicated in decision making in both humans^{15,19} and primates^{24–26}, with different subregions being associated with different output modalities. In lateral intraparietal area LIP, associated with saccades, neurons also carry information about decision variables such as the reward expected for a saccade^{24–26}; the area perhaps serves as an interface between frontal areas (where such information may be calculated) and motor output. The anterior border of the sulcus, close to our exploration-related activation, is associated with grasping and manual manipulation²⁷, raising the possibility that such information (here, that associated with exploration) might also reach parietal regions involved in the button-press actions in our task.

Last, we used a multiple regression analysis to verify that differential activity in frontopolar and intraparietal regions during exploratory trials was not better explained by any of several potentially confounding factors such as switching between options or reaction times (see Supplementary Information and Supplementary Tables 4 and 5).

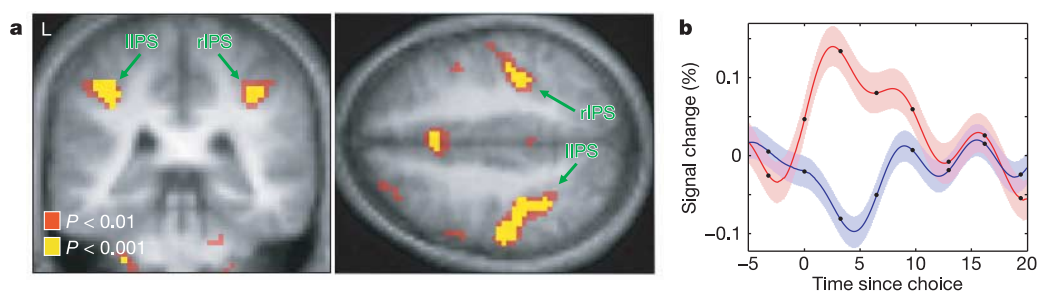


Figure 4 | Exploration-related activity in intraparietal sulcus. **a**, Regions of left and right intraparietal sulcus (lIPS and rIPS) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a subject-averaged structural scan. The coordinates of the activated areas are $[-29, -33, 45, \text{peak } z = 4.39]$ for lIPS and $[39, -36, 42, \text{peak } z = 4.16]$ for

rIPS. **b**, lIPS BOLD time courses averaged over 1,515 exploratory (red line) and 2,646 exploitative (blue line) decisions. Black dots indicate the sampling frequency (although, because sample alignment varied from trial to trial, time courses were upsampled). Coloured fringes show error bars (representing s.e.m.).

These results have important implications for both computational and neural accounts of action selection. The finding of brain regions discretely implicated in exploration (and particularly that one of them is a prefrontal, high-level control structure²¹) is consistent with a theory in which exploration is accomplished by overriding an exploitative tendency, but troubling for accounts such as uncertainty bonus schemes^{1,14}, which more tightly entangle exploration and exploitation. Such anatomical separation would be unlikely under these latter schemes, because they work by choosing actions with respect to a unified value metric that simultaneously prizes both information gathering and primary reward. Just such an exploration-encouraging value metric has previously been suggested to explain why dopamine neurons respond to novel, neutral stimuli¹³; such anomalous responses in an otherwise typically appetitive signal remain puzzling in view of our failure here to find either behavioural or neural evidence for such an account.

Exploration has a central role in the acquisition of adaptive behaviour in environments that change. Characteristic expressions of frontal pathology²⁸ include impairments in task switching as well as behavioural perseveration, which might relate, at least in part, to a core deficit in exploration. As one might expect for such a critical function, subcortical systems are also implicated in the control of exploration, with noradrenaline being suggested as regulating a global propensity to explore^{29,30}, a factor captured in our model in terms of the parameter regulating competition in the softmax rule. Last, self-directed exploration of the form studied here is an example of a refined cognitive function that is ubiquitous but hard to pin down in regular designs (because exploratory and exploitative responses are apparently seamlessly mixed). We were able to capture it only through a tight coupling of computational modelling, behavioural analysis and functional neuroimaging.

METHODS

Fourteen right-handed healthy human subjects participated in an fMRI scan (using a 1.5 T Siemens Sonata scanner) while repeatedly choosing between animated slot machines. One of three candidate reinforcement learning models for their behaviour was selected, and its parameters estimated, by maximizing the cumulative likelihood of the subjects' choices given the model and parameters. Trials were classified according to the model as exploratory or exploitative, and trial-by-trial estimates of subjects' predictions about slot machine payoffs (and the error or mismatch between those predictions and received payoffs) were generated by running the model progressively on the subjects' actual choices and winnings. A general linear model implemented in SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, UCL) was used to locate brain voxels where the measured BOLD signal was significantly correlated with these model-generated signals. Regions identified as significantly correlated with exploration were subjected to a subsequent multiple regression analysis to investigate whether other, confounding factors might better account for the observed activity. For a detailed description of the experimental and analytical techniques, see Supplementary Methods.

Received 7 February; accepted 30 March 2006.

- Gittins, J. C. & Jones, D. in *Progress in Statistics* (ed. Gani, J.) 241–266 (North-Holland, Amsterdam, 1974).
- Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
- Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
- Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
- Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C. & Fiez, J. A. Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* **84**, 3072–3077 (2000).
- Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. fMRI visualization of

- brain activity during a monetary incentive delay task. *Neuroimage* **12**, 20–27 (2000).
- McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
- O'Doherty, J. P. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
- Charnov, E. L. Optimal foraging: The marginal value theorem. *Theor. Popul. Biol.* **9**, 129–136 (1976).
- Owen, A. M. Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog. Neurobiol.* **53**, 431–450 (1997).
- Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control. *Nature Neurosci.* **8**, 1704–1711 (2005).
- Kakade, S. & Dayan, P. Dopamine: Generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).
- Kaelbling, L. P. *Learning in Embedded Systems* (MIT Press, Cambridge, Massachusetts, 1993).
- McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Separate neural systems value immediate and delayed monetary rewards. *Science* **306**, 503–507 (2004).
- O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neurosci.* **4**, 95–102 (2001).
- O'Doherty, J. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr. Opin. Neurobiol.* **14**, 769–776 (2004).
- Gottfried, J. A., O'Doherty, J. & Dolan, R. J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104–1107 (2003).
- Tanaka, S. C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neurosci.* **7**, 887–893 (2004).
- Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
- Ramnani, N. & Owen, A. M. Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Rev. Neurosci.* **5**, 184–194 (2004).
- Koechlin, E., Ody, C. & Kouneiher, F. A. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).
- Braver, T. S. & Bongiolatti, S. R. The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage* **15**, 523–536 (2002).
- Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
- Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behaviour and the representation of value in the parietal cortex. *Science* **304**, 1782–1787 (2004).
- Dorris, M. C. & Glimcher, P. W. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44**, 365–378 (2004).
- Grefkes, C. & Fink, G. R. The functional organization of the intraparietal sulcus in humans and monkeys. *J. Anat.* **207**, 3–17 (2005).
- Burgess, P. W., Veitch, E., de Lacy Costello, A. & Shallice, T. The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia* **38**, 848–863 (2000).
- Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554 (1999).
- Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495–506 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Li, S. McClure, B. King-Casas and P. R. Montague for sharing their unpublished data on exploration, and Y. Niv, Z. Gharamani and C. Camerer for discussions. Funding was from a Royal Society USA Research Fellowship (N.D.), the Gatsby Foundation (N.D., P.D.), the EU BIBA project (N.D., P.D.), and a Wellcome Trust Programme Grant (J.O.D., R.D.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.D. (daw@gatsby.ucl.ac.uk) or J.O.D. (jdoherty@hss.caltech.edu).

LETTERS

Loss of autophagy in the central nervous system causes neurodegeneration in mice

Masaaki Komatsu^{1,2*}, Satoshi Waguri^{3*†}, Tomoki Chiba¹, Shigeo Murata¹, Jun-ichi Iwata^{1,2}, Isei Tanida², Takashi Ueno², Masato Koike³, Yasuo Uchiyama³, Eiki Kominami² & Keiji Tanaka¹

Protein quality-control, especially the removal of proteins with aberrant structures, has an important role in maintaining the homeostasis of non-dividing neural cells¹. In addition to the ubiquitin–proteasome system, emerging evidence points to the importance of autophagy—the bulk protein degradation pathway involved in starvation-induced and constitutive protein turnover—in the protein quality-control process^{2,3}. However, little is known about the precise roles of autophagy in neurons. Here we report that loss of *Atg7* (autophagy-related 7), a gene essential for autophagy, leads to neurodegeneration. We found that mice lacking *Atg7* specifically in the central nervous system showed behavioural defects, including abnormal limb-clasping reflexes and a reduction in coordinated movement, and died within 28 weeks of birth. *Atg7* deficiency caused massive neuronal loss in the cerebral and cerebellar cortices. Notably, polyubiquitinated proteins accumulated in autophagy-deficient neurons as inclusion bodies, which increased in size and number with ageing. There was, however, no obvious alteration in proteasome function. Our results indicate that autophagy is essential for the survival of neural cells, and that impairment of autophagy is implicated in the pathogenesis of neurodegenerative disorders involving ubiquitin-containing inclusion bodies.

Macroautophagy (hereafter referred to as autophagy) is an evolutionarily conserved pathway in which the cytoplasm and organelles are engulfed within double-membraned vesicles, known as autophagosomes, in preparation for the turnover and recycling of these cellular constituents⁴. Genetic studies using various model organisms have highlighted the importance of autophagy in physiological and pathological events⁵. The principal role of autophagy is in the supply of nutrients for survival, as shown in yeast⁶ and early neonatal mice^{7,8}. Autophagy also has a role in cellular remodelling during differentiation and the development of multicellular organisms, such as dauer formation in *Caenorhabditis elegans*⁹ and metamorphosis in *Drosophila melanogaster*¹⁰. Moreover, constitutive autophagy, which occurs independently of nutrient stress, contributes to mouse liver homeostasis⁸, major histocompatibility class (MHC) II antigen presentation¹¹, and cellular defence against invading streptococci¹² and *Mycobacterium tuberculosis*¹³. However, the physiological functions of autophagy, particularly in neurons, are still largely unknown.

To examine the relationship between neuronal pathology and autophagy deficiency *in vivo*, we crossed *Atg7*-conditional knockout mice (*Atg7^{flox/flox}*) (ref. 8) with transgenic mice expressing Cre recombinase under the control of the nestin promoter (*nestin-Cre*) (ref. 14), to produce mice deficient for *Atg7* specifically in the central nervous system (*Atg7^{flox/flox}; nestin-Cre*). *Atg7* is an E1-like enzyme for both the *Atg12*- and *Atg8*-conjugation systems¹⁵, and is essential

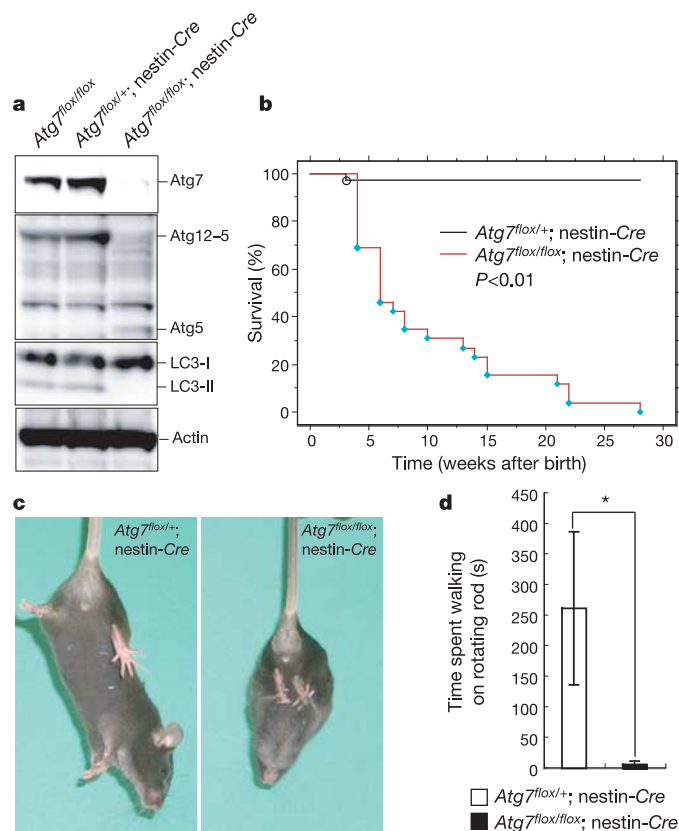


Figure 1 | Behavioural disorder in *Atg7^{flox/flox}; nestin-Cre* mice.

a, Impairment of two ATG-conjugation systems (*Atg12* and *LC3*) in the *Atg7*-deficient brain. Brain homogenates from P28 mice were immunoblotted with antibodies against *Atg7*, *Atg5* and *LC3*. Actin was used as a loading control. Data shown are representative of three separate experiments. **b**, Kaplan–Meier survival curves of *Atg7^{flox/+}; nestin-Cre* ($n = 41$) and *Atg7^{flox/flox}; nestin-Cre* ($n = 26$) mice over 28 weeks. **c**, Abnormal limb-clasping reflexes in *Atg7^{flox/flox}; nestin-Cre* mice at P28. When lifted by the tail, *Atg7^{flox/+}; nestin-Cre* mice behave normally, extending their hind limbs and bodies. In contrast, *Atg7^{flox/flox}; nestin-Cre* mice bend their legs towards their trunk or tighten their back limbs to their bodies and anterior limbs. **d**, Movement ataxia in *Atg7^{flox/flox}; nestin-Cre* mice at P28. Motor coordination was tested using a rotarod assay. *Atg7^{flox/+}; nestin-Cre* ($n = 5$) and *Atg7^{flox/flox}; nestin-Cre* ($n = 5$) mice were placed on a rod rotating at 20 r.p.m., and the time spent on the rod was counted. Data show mean $\pm$ s.d. *, $P < 0.01$ (Student's *t*-test). There was no significant sex difference in survival rate and onset-stage of abnormal limb-clasping and tremor in *Atg7^{flox/flox}; nestin-Cre* mice.

¹Laboratory of Frontier Science, Tokyo Metropolitan Institute of Medical Science, Bunkyo-ku, Tokyo 113-8613, Japan. ²Department of Biochemistry, Juntendo University School of Medicine, Bunkyo-ku, Tokyo 113-8421, Japan. ³Department of Cell Biology and Neurosciences, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan. [†]Present address: Department of Anatomy and Histology, Fukushima Medical University School of Medicine, 1 Hikarigaoka, Fukushima 960-1295, Japan.

*These authors contributed equally to this work.

for autophagy⁸. Atg7 protein was absent at postnatal day (P)28 in brain from *Atg7^{fllox/fllox}*; nestin-Cre but not control (*Atg7^{fllox/+}*; nestin-Cre) mice (Fig. 1a). The level of Atg7 protein in other tissues such as liver, lung, heart and muscle was comparable between *Atg7^{fllox/fllox}*; nestin-Cre and control mice (data not shown). Atg12–Atg5 conjugate was detected only in the brains of control mice by immunoblotting with an anti-Atg5 antibody (Fig. 1a). In contrast, free Atg5, which was faintly observed in the control mouse brain, was clearly increased in the mutant brain (Fig. 1a). The mammalian homologue of yeast Atg8, microtubule-associated protein 1 light-chain 3 (LC3), exists in two forms (LC3-I and LC3-II)¹⁶. Both forms were detected in brains from control mice, but only the LC3-I form was detected in *Atg7^{fllox/fllox}*; nestin-Cre brain (Fig. 1a). The loss of both Atg7 and LC3-II proteins was observed from P0 in the brain of *Atg7^{fllox/fllox}*; nestin-Cre mice (Supplementary Fig. S1). These results indicate complete impairment of autophagy in the central nervous system of *Atg7^{fllox/fllox}*; nestin-Cre mice after birth.

Atg7^{fllox/fllox}; nestin-Cre mice were viable at birth and indistinguishable in appearance from their littermates. However, the survival rate of the mutant mice diminished markedly by four weeks after birth, and all *Atg7^{fllox/fllox}*; nestin-Cre mice were dead within 28 weeks

(Fig. 1b). We also observed growth retardation as early as P14 in these mice (data not shown). Furthermore, the mice showed motor and behavioural deficits, including abnormal limb-clasping reflexes (Fig. 1c) and tremor, and in some cases, they walked on their tiptoes. In a rotarod test, most *Atg7^{fllox/fllox}*; nestin-Cre mice fell after grasping the rod only briefly (Fig. 1d). These motor and behavioural deficits began to appear at P14–P21. These results suggest that autophagy deficiency in the central nervous system results in a severe neurological disorder.

Histological analysis using Meyer's haematoxylin and eosin (H&E) staining showed marked atrophy of the cerebral cortical region of *Atg7^{fllox/fllox}*; nestin-Cre brain at P56 (Fig. 2a, b). The ratios of cortical thickness to dorsoventral thickness of the brain in *Atg7^{fllox/+}*; nestin-Cre and *Atg7^{fllox/fllox}*; nestin-Cre mice were 0.17 ± 0.00031 and 0.15 ± 0.00034 , respectively ($n = 5$, $P < 0.01$). Notably, almost no large pyramidal neurons were observed in *Atg7^{fllox/fllox}*; nestin-Cre mice compared with the corresponding region in brains from control mice (Fig. 2c, d). Immunostaining for the glial marker GFAP (glial fibrillary acidic protein) showed an increase in GFAP signal in the cerebral cortex of *Atg7^{fllox/fllox}*; nestin-Cre mice (Fig. 2e, f), suggesting the presence of neuronal damage in this region. In the cerebellar

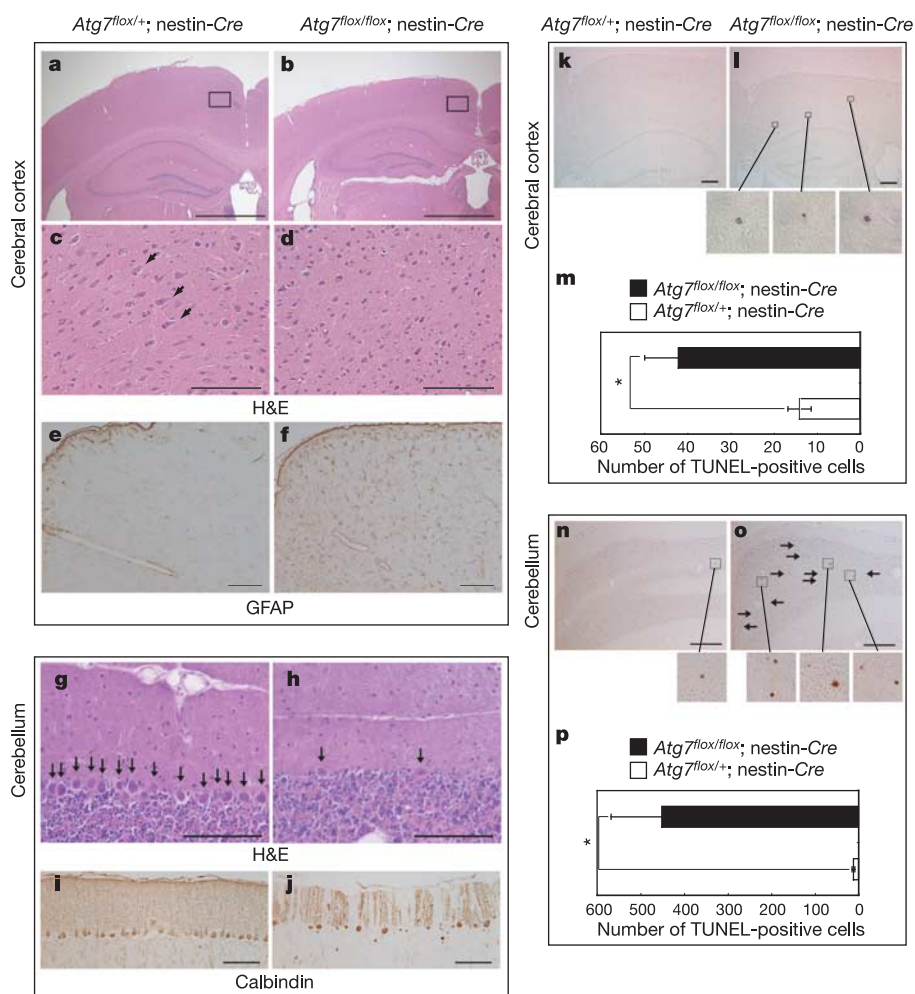


Figure 2 | Marked cell death in autophagy-deficient cerebral cortex and cerebellum. **a–f**, Histological analyses of *Atg7^{fllox/+}*; nestin-Cre (left) and *Atg7^{fllox/fllox}*; nestin-Cre (right) cerebral cortex at P56. Cryosections were stained with H&E (**a–d**) or immunostained for the glial marker GFAP (**e, f**). Boxed areas in **a** and **b** are magnified in **c** and **d**, respectively. Arrows in **c** point to large pyramidal neurons in the cerebral cortex. **g–j**, Histological analysis of *Atg7^{fllox/+}*; nestin-Cre (left) and *Atg7^{fllox/fllox}*; nestin-Cre (right) cerebellum at P56. Cryosections were stained with H&E (**g, h**) or immunostained for the Purkinje marker calbindin (**i, j**). Arrows in **g** and **h**

indicate cerebellar Purkinje cells. **k–p**, TUNEL staining of the cerebral cortex (**k, l**) and cerebellum (**n, o**) at P56 in *Atg7^{fllox/+}*; nestin-Cre (**k, n**) and *Atg7^{fllox/fllox}*; nestin-Cre (**l, o**) sections. TUNEL-positive cells are indicated with arrows and shown as higher magnification images beneath panels **l, n** and **o**. Histograms show the average number ($\pm$ s.d.) of TUNEL-positive cells in ten sections for three animals of each genotype (**m, p**). *, $P < 0.05$ (t -test). Scale bars, 1 mm (**a, b**), 100 μ m (**c–j**), 250 μ m (**k, l, n, o**). We observed no sex difference in brain morphology or neuronal loss in *Atg7^{fllox/fllox}*; nestin-Cre mice.

cortex, H&E staining revealed a large reduction in the number of Purkinje cells in the mutant brain (Fig. 2g, h), which was further confirmed by immunolabelling of Purkinje cells with an anti-calbindin antibody (Fig. 2i, j). Similar neuronal loss was also recognized in the hippocampal pyramidal cell layer of mutant brain (Supplementary Fig. S2).

To determine whether the reduced number of neurons observed in *Atg7^{flox/flox}; nestin-Cre* mouse brain was caused by cell death, we performed TUNEL (TdT-mediated dUTP nick end labelling) assays. We observed a marked increase in the number of TUNEL-positive cells in the cerebral cortex (Fig. 2k–m) and granular cell layer of the cerebellum at P56 in *Atg7^{flox/flox}; nestin-Cre* mice (Fig. 2n–p) compared with control mouse brains. Although loss of Purkinje cells was observed in the mutant brain, we could not detect TUNEL-positive Purkinje cells at any developmental stage examined. However, when *Atg7* was specifically depleted in Purkinje cells using transgenic mice expressing Cre recombinase under the control of the *Pcp2* gene

promoter (*Pcp2-Cre*), a marked reduction in the number of Purkinje cells was detected in the absence of TUNEL reactivity (data not shown), suggesting that neurons deficient in autophagy can die in a cell-autonomous fashion. Together, these results indicate that lack of autophagy in the central nervous system leads to neurodegeneration.

We have previously reported that autophagy is responsible for constitutive protein turnover in quiescent hepatocytes even under nutrient-rich conditions, and that a defect in autophagy leads to the accumulation of large, ubiquitin-containing inclusion bodies⁸. We therefore probed brain sections with an anti-ubiquitin antibody to examine the presence of ubiquitin-containing inclusion bodies. At P56, ubiquitin-positive dots were detected in several regions of the *Atg7^{flox/flox}; nestin-Cre* mouse brain, including the cerebral cortex (Fig. 3b), cerebellar Purkinje cells (Fig. 3d), hippocampal pyramidal neurons (Fig. 3f), thalamus (data not shown), hypothalamus (Fig. 3h), amygdala (Fig. 3j) and pontine nuclei (Fig. 3l). The degree of staining varied by region. For example, whereas most neuronal

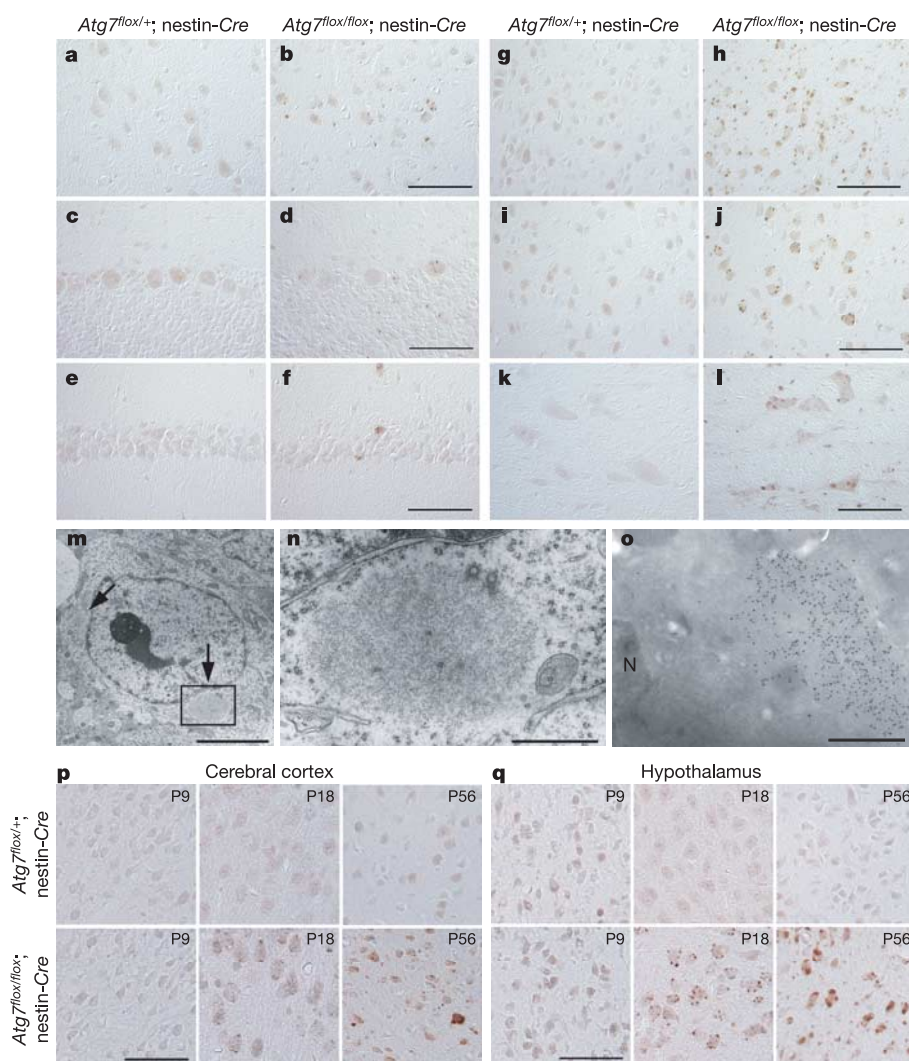


Figure 3 | Appearance of ubiquitin-positive inclusions in autophagy-deficient neurons. **a–l**, The presence of ubiquitin-positive dots was examined immunohistochemically in several regions including cerebral cortex (**a, b**), cerebellum (**c, d**), hippocampus (**e, f**), hypothalamus (**g, h**), amygdala (**i, j**) and pontine nuclei (**k, l**) of *Atg7^{flox/+}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* mice. Note the presence of numerous ubiquitin dots in the amygdala and hypothalamus of the representative mutants. Scale bars, 50 μ m. **m, n**, Electron micrographs of the brain of *Atg7^{flox/flox}; nestin-Cre* mice. Inclusion bodies (arrows) were often observed in *Atg7^{flox/flox}; nestin-Cre* hypothalamus. The boxed region in **m** is shown in **n**. Inclusion bodies

were not detected in *Atg7^{flox/+}; nestin-Cre* brain (data not shown). Scale bars, 5 μ m (**m**), 1 μ m (**n**). **o**, Immunoelectron micrograph of ubiquitin in a representative *Atg7^{flox/flox}; nestin-Cre* hypothalamus. N, nucleus. Scale bar, 1 μ m. **p, q**, Immunohistochemical detection of ubiquitin-positive inclusions in the cerebral cortex (**p**) and hypothalamus (**q**) at P9, P18 and P56. Brain sections of each genotype at the indicated ages were immunostained with an anti-ubiquitin antibody. Ubiquitin-positive inclusions appeared at P18 and became larger with ageing in the brain of *Atg7^{flox/flox}; nestin-Cre* mice. Scale bars, 50 μ m.

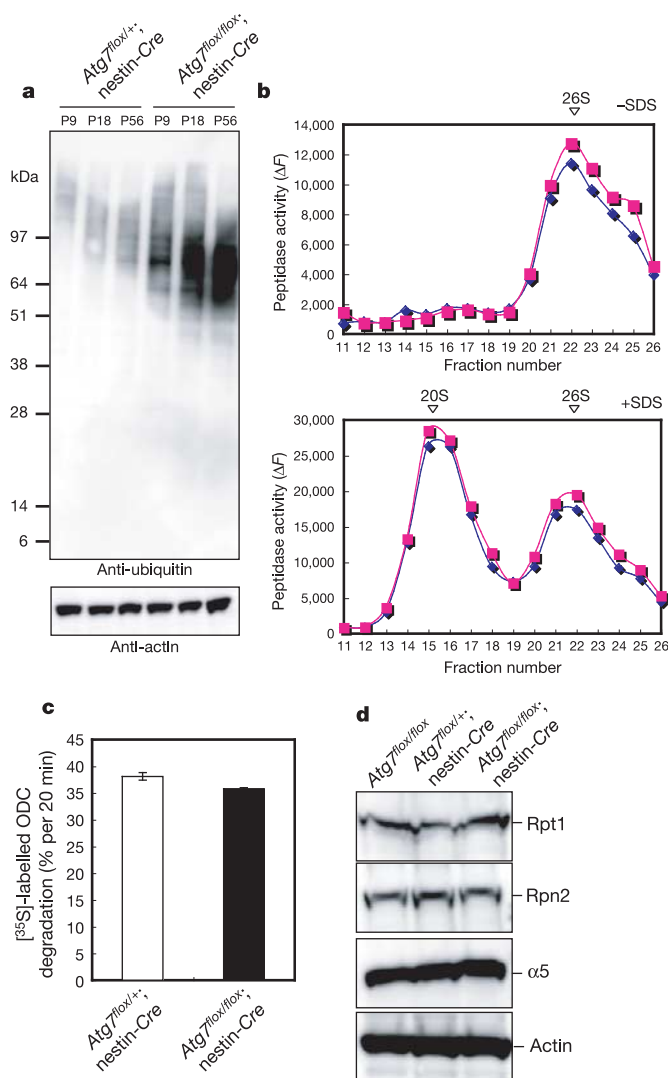


Figure 4 | Qualitative and quantitative analyses of proteasomes in autophagy-deficient brain. **a**, Increase in ubiquitinated proteins in *Atg7^{flox/flox}; nestin-Cre* mouse brain over time. Homogenates of P9, P18 and P56 brains from *Atg7^{flox/flox}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* mice were immunoblotted with an anti-ubiquitin antibody. An anti-actin antibody was used as a loading control. Data shown are representative of three separate experiments. We observed no sex difference in the accumulation of ubiquitin in *Atg7^{flox/flox}; nestin-Cre* mice. **b**, Peptide hydrolysis activity of 20S and 26S proteasomes. Homogenates from P28 *Atg7^{flox/flox}; nestin-Cre* (blue) and *Atg7^{flox/flox}; nestin-Cre* (pink) brains were fractionated by glycerol density gradient centrifugation (10–40% glycerol from fraction 1 to fraction 30). Aliquots from each fraction were used for the assay of chymotryptic activity of proteasomes using Suc-LLVY-AMC as a substrate in the absence (top) or presence (bottom) of 0.05% SDS. The sedimenting positions of 20S and 26S proteasomes are indicated with arrowheads. Note that whereas 26S proteasomes exist in active forms in tissues, 20S proteasomes are latent and are activated artificially by a low concentration of SDS. **c**, ATP-dependent degradation of [³⁵S]-labelled ODC. Degradation of [³⁵S]-labelled ODC was assayed using crude extracts from P28 *Atg7^{flox/flox}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* brains. The experiment was repeated three times, and values represent mean $\pm$ s.d. In the above assays, there were no significant differences between *Atg7^{flox/flox}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* mice. **d**, Immunoblot analysis of 26S proteasome components. Homogenates from P28 *Atg7^{flox/flox}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* brains were immunoblotted with antibodies against the indicated proteins. Data shown are representative of three separate experiments. There was no change in proteasome status for the different genotypes.

cells in the amygdala (Fig. 3j) and hypothalamus (Fig. 3h) contained several ubiquitin dots of small to large size, only a small number of cerebellar Purkinje cells stained for ubiquitin, and the immunoreactive dots were of small size (Fig. 3d). Electron microscopy showed that *Atg7^{flox/flox}; nestin-Cre* hypothalamic neurons had circular or elliptical large structures composed of fibrillar elements in the perikarya (Fig. 3m, n). Immunoelectron microscopy further confirmed that these aberrant structures contained ubiquitin (Fig. 3o). These ubiquitin dots appeared not only in the perikarya of neurons, but also in the intercellular space (see Fig. 3h, j, l). Dots found in the intercellular space might correspond to ubiquitin inside neurites, because they were observed in myelinated axons around pontine nuclei using both light and electron microscopy (Supplementary Fig. S3a, b). In contrast, almost no ubiquitin dots were observed in astroglial cells (Supplementary Fig. S3c). Together with the results in Fig. 3, we concluded that most of the distinct ubiquitin dots were located in neurons.

We examined the development of ubiquitin-containing inclusion bodies at different stages by dissecting the brains of *Atg7^{flox/flox}; nestin-Cre* and *Atg7^{flox/flox}; nestin-Cre* mice at P9, P18 and P56. Only a few ubiquitin-positive aggregates were noted in the neurons of control and mutant cerebral cortex and hypothalamus at P9 (Fig. 3p, q). In contrast, several ubiquitin-containing inclusions were clearly noted in *Atg7^{flox/flox}; nestin-Cre* cortex and hypothalamus at P18, increasing in number and size by P56 (Fig. 3p, q). These results indicate an age-dependent increase in ubiquitin-containing inclusion bodies in autophagy-deficient neurons. Consistent with the above immunohistochemical analysis (Fig. 3p, q), immunoblot analysis revealed increasing levels of high-molecular-mass polyubiquitinated proteins with age in the brains of *Atg7^{flox/flox}; nestin-Cre* mice (Fig. 4a), and an increase in their insoluble forms at later developmental stages (data not shown).

Finally, we examined whether autophagy deficiency influences proteasome functions. The chymotryptic activities of 26S and 20S proteasomes (measured using Suc-LLVY-MCA as a substrate) were comparable in extracts from both control and *Atg7^{flox/flox}; nestin-Cre* brains (Fig. 4b). Furthermore, the ATP-dependent degradation of ornithine decarboxylase (ODC) by 26S proteasomes was similar in control and mutant brains (Fig. 4c). Moreover, the relative amounts of several subunits of the 26S proteasome did not change in the brain irrespective of autophagy deficiency (as detected by immunoblotting, Fig. 4d). These results indicate that age-dependent accumulation of ubiquitin-positive aggregates in the autophagy-deficient brain occurs despite the apparently normal function of proteasomes.

Over the past decade, researchers working in the field of neurodegenerative diseases have made great progress in uncovering the mechanisms of these disorders by focusing on the interplay between proteolytic stress and neural cell death^{17,18}. Increasing evidence indicates that ubiquitin-positive inclusion bodies—the pathological hallmark of various neurodegenerative diseases—are formed by dysfunction of proteasome degrading machinery³. Indeed, proteins with aberrant structure impair proteasome functions directly, thus attenuating ‘garbage disposal’¹⁹. On the other hand, the accumulation of autophagosomes owing to impairment of fusion with lysosomes is observed in various disorders, including Alzheimer’s disease^{20–22}, and it has been proposed that autophagy functions to degrade toxic proteins in familial neurodegenerative diseases^{23–25}. However, it remains unknown whether these two proteolytic systems work independently or cooperatively to maintain protein homeostasis in cells. Furthermore, whether autophagy has a role in cell death or cell survival is currently under debate²⁶.

We have shown that *Atg7^{flox/flox}; nestin-Cre* mice exhibit neurological abnormalities and neuronal death, suggesting that impaired autophagy causes neurodegeneration. We suggest a particularly important role for autophagy in the brain, to which nutrients must be constantly supplied from other organs, even under fasting conditions. Moreover, we find that autophagy deficiency in neurons leads to the accumulation of ubiquitin-containing inclusion bodies,

without obvious deficits in proteasome function. Hence, our data indicate a central role for constitutive autophagy in the elimination of unfavourable proteins and in the survival of neurons, independent of the proteasome system (see the proposed model in Supplementary Fig. S4). Although we do not know whether autophagy and proteasome degradation target a similar set of normal and/or misfolded proteins, it is plausible that the autophagic pathway assists in degrading accumulated intractable proteins when cellular levels of aberrant proteins overwhelm the disposal capacity of the proteasome.

We have shown that a lack of autophagy is associated with neurodegeneration, even in the absence of harmful gene products found in neurodegenerative disorders such as Huntington's disease, Parkinson's disease and amyotrophic lateral sclerosis. We therefore predict that the role of autophagy becomes even more critical in the pathogenesis of such neurodegenerative diseases, when disease-related, aggregation-prone proteins are expressed as a result of genetic mutations and/or environmental insults, leading to early-onset symptoms.

METHODS

Animals. Nestin-Cre transgenic mice¹⁴ were purchased from the Jackson Laboratory. *Atg7^{flax/flax}* mice⁸ were bred with nestin-Cre transgenic mice to produce *Atg7^{flax/flax}*; nestin-Cre mice. Mice were housed in a pathogen-free facility. Motor function was assessed using a rotarod test²⁷. Experimental protocols were approved by the Ethics Review Committee for Animal Experimentation at the Tokyo Metropolitan Institute of Medical Science.

Immunoblot analysis. Immunoblots were carried out as described previously⁸. Antibodies against Atg7, Atg5 and LC3 have been described previously⁸. Antibodies against Rpt1, Rpn2 and $\alpha 5$ were provided by K. B. Hendil. Polyclonal anti-ubiquitin (FK2; Medical & Biological Laboratories) and anti-actin (MAB1501R; Chemicon) antibodies were also used.

Histological examination. *Atg7^{flax/+}*; nestin-Cre and *Atg7^{flax/flax}*; nestin-Cre mice were fixed by cardiac perfusion with 0.1 M phosphate buffer containing 4% paraformaldehyde, 4% sucrose for light microscopy and immunohistochemistry, with 0.1 M phosphate buffer containing 2% paraformaldehyde, 2% glutaraldehyde for standard electron microscopy, or with 0.1 M phosphate buffer containing 4% paraformaldehyde, 0.1% glutaraldehyde for immunoelectron microscopy. Brain tissues were excised and processed for morphological analysis as described previously^{8,28}. For light microscopic analysis, 10- μ m cryosections were cut and stained with H&E or immunolabelled with the following antibodies: anti-human NeuN (Abcam), anti-GFAP (Sigma), anti-calbindin (Sigma), anti-myelin basic protein (MBP; MCA409S, Serotec) and anti-ubiquitin (DAKO) antibodies. The TUNEL assay has been described previously²⁸. For counting TUNEL-positive signals in the cerebral cortex, 60 coronal sections containing the anterior portion of the hippocampus (~0.6-mm thick in total) were cut, and TUNEL staining was performed on every sixth section, on a total of ten sections.

Electron microscopy and immunoelectron microscopy. Fixed brains were post-fixed with 1% OsO₄, embedded in Epon812 and sectioned. Immunoelectron microscopy was carried out on cryothin sections as described previously²⁹. In brief, brains were frozen in phosphate buffer containing 2.3 M sucrose and 20% polyvinyl pyrrolidone. Ultrathin sections were mounted on Formvar carbon-coated nickel grids, blocked with 1% bovine serum albumin (BSA) in PBS, and incubated with anti-ubiquitin antibody (1B3) and colloidal gold-conjugated secondary antibody.

Glycerol gradient analysis. Samples were fractionated by 10–40% (v/v) linear glycerol density gradient centrifugation (22 h, 100,000g) as described previously³⁰.

Assay of proteasome activity. Peptidase activity was measured using a fluorescent peptide substrate, succinyl-Leu-Leu-Val-Tyr-7-amido-4-methylcoumarin (Suc-LLVY-MCA), as described previously³⁰. Ornithine decarboxylase (ODC)-degradation activity was assayed as described previously³⁰.

Received 6 February; accepted 20 March 2006.

Published online 19 April 2006.

- Forman, M. S., Trojanowski, J. Q. & Lee, V. M. Neurodegenerative diseases: a decade of discoveries paves the way for therapeutic breakthroughs. *Nature Med.* **10**, 1055–1063 (2004).
- Shintani, T. & Klionsky, D. J. Autophagy in health and disease: a double-edged sword. *Science* **306**, 990–995 (2004).
- Goldberg, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895–899 (2003).
- Reggiori, F. & Klionsky, D. J. Autophagosomes: biogenesis from scratch? *Curr. Opin. Cell Biol.* **17**, 415–422 (2005).

- Levine, B. & Klionsky, D. J. Development by self-digestion: molecular mechanisms and biological functions of autophagy. *Dev. Cell* **6**, 463–477 (2004).
- Tsukada, M. & Ohsumi, Y. Isolation and characterization of autophagy-defective mutants of *Saccharomyces cerevisiae*. *FEBS Lett.* **333**, 169–174 (1993).
- Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432**, 1032–1036 (2004).
- Komatsu, M. *et al.* Impairment of starvation-induced and constitutive autophagy in *Atg7*-deficient mice. *J. Cell Biol.* **169**, 425–434 (2005).
- Melendez, A. *et al.* Autophagy genes are essential for dauer development and life-span extension in *C. elegans*. *Science* **301**, 1387–1391 (2003).
- Juhász, G., Csikos, G., Sinka, R., Erdelyi, M. & Sass, M. The *Drosophila* homolog of Atg1 is essential for autophagy and development. *FEBS Lett.* **543**, 154–158 (2003).
- Paludan, C. *et al.* Endogenous MHC class II processing of a viral nuclear antigen after autophagy. *Science* **307**, 593–596 (2005).
- Nakagawa, I. *et al.* Autophagy defends cells against invading group A *Streptococcus*. *Science* **306**, 1037–1040 (2004).
- Gutierrez, M. G. *et al.* Autophagy is a defense mechanism inhibiting BCG and *Mycobacterium tuberculosis* survival in infected macrophages. *Cell* **119**, 753–766 (2004).
- Betz, U. A., Vosschenrich, C. A., Rajewsky, K. & Muller, W. Bypass of lethality with mosaic mice generated by Cre-loxP-mediated recombination. *Curr. Biol.* **6**, 1307–1316 (1996).
- Ohsumi, Y. Molecular dissection of autophagy: two ubiquitin-like systems. *Nature Rev. Mol. Cell Biol.* **2**, 211–216 (2001).
- Kabeya, Y. *et al.* LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosomal membranes after processing. *EMBO J.* **19**, 5720–5728 (2000).
- Ciechanover, A. & Brundin, P. The ubiquitin proteasome system in neurodegenerative diseases: sometimes the chicken, sometimes the egg. *Neuron* **40**, 427–446 (2003).
- Bossy-Wetzel, E., Schwarzenbacher, R. & Lipton, S. A. Molecular pathways to neurodegeneration. *Nature Med.* **10** (Suppl.), S2–S9 (2004).
- Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* **292**, 1552–1555 (2001).
- Yu, W. H. *et al.* Macroautophagy—a novel β -amyloid peptide-generating pathway activated in Alzheimer's disease. *J. Cell Biol.* **171**, 87–98 (2005).
- Tanaka, Y. *et al.* Accumulation of autophagic vacuoles and cardiomyopathy in LAMP-2-deficient mice. *Nature* **406**, 902–906 (2000).
- Nishino, I. *et al.* Primary LAMP-2 deficiency causes X-linked vacuolar cardiomyopathy and myopathy (Danon disease). *Nature* **406**, 906–910 (2000).
- Webb, J. L., Ravikumar, B., Atkins, J., Skepper, J. N. & Rubinsztein, D. C. α -Synuclein is degraded by both autophagy and the proteasome. *J. Biol. Chem.* **278**, 25009–25013 (2003).
- Ravikumar, B. *et al.* Inhibition of mTOR induces autophagy and reduces toxicity of polyglutamine expansions in fly and mouse models of Huntington disease. *Nature Genet.* **36**, 585–595 (2004).
- Fortun, J., Dunn, W. A. Jr, Joy, S., Li, J. & Notterpek, L. Emerging role for autophagy in the removal of aggregates in Schwann cells. *J. Neurosci.* **23**, 10672–10680 (2003).
- Levine, B. & Yuan, J. Autophagy in cell death: an innocent convict? *J. Clin. Invest.* **115**, 2679–2688 (2005).
- Bontekoe, C. J. *et al.* Knockout mouse model for *Fxr2*: a model for mental retardation. *Hum. Mol. Genet.* **11**, 487–498 (2002).
- Koike, M. *et al.* Involvement of two different cell death pathways in retinal atrophy of cathepsin D-deficient mice. *Mol. Cell. Neurosci.* **22**, 146–161 (2003).
- Waguri, S. *et al.* Cysteine proteinases in GH4C1 cells, a rat pituitary tumour cell line, are secreted by the constitutive and regulated secretory pathways. *Eur. J. Cell Biol.* **67**, 308–318 (1995).
- Tanahashi, N. *et al.* Hybrid proteasomes. Induction by interferon- γ and contribution to ATP-dependent proteolysis. *J. Biol. Chem.* **275**, 14336–14345 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Kaneko, T. Kouno and K. Tatsumi for technical assistance. We also thank A. Yabashi, K. Kanno, F. Kaji and K. Ikeue for help with morphological analysis, J. Ezaki for discussion, and Z. Yue for critical reading of the manuscript. This work was supported in part by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author contributions M.K. and T.C. generated *Atg7^{flax/flax}* mice, and M.K. and J.I. performed most of the experiments to characterize *Atg7^{flax/flax}*; nestin-Cre mice. S.W. performed histological and microscopic analyses, and S.M. performed the biochemical analysis of proteasome activity. M.K., S.W. and K.T. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.T. (tanakak@rinshoken.or.jp).

Suppression of basal autophagy in neural cells causes neurodegenerative disease in mice

Taichi Hara¹, Kenji Nakamura², Makoto Matsui^{1,3,4}, Akitsugu Yamamoto⁵, Yohko Nakahara², Rika Suzuki-Migishima², Minesuke Yokoyama⁶, Kenji Mishima⁷, Ichiro Saito⁷, Hideyuki Okano^{8,9} & Noboru Mizushima^{1,10}

Autophagy is an intracellular bulk degradation process through which a portion of the cytoplasm is delivered to lysosomes to be degraded^{1–4}. Although the primary role of autophagy in many organisms is in adaptation to starvation, autophagy is also thought to be important for normal turnover of cytoplasmic contents, particularly in quiescent cells such as neurons. Autophagy may have a protective role against the development of a number of neurodegenerative diseases^{5–8}. Here we report that loss of autophagy causes neurodegeneration even in the absence of any disease-associated mutant proteins. Mice deficient for *Atg5* (autophagy-related 5) specifically in neural cells develop progressive deficits in motor function that are accompanied by the accumulation of cytoplasmic inclusion bodies in neurons. In *Atg5*^{−/−} cells, diffuse, abnormal intracellular proteins accumulate, and then form aggregates and inclusions. These results suggest that the continuous clearance of diffuse cytosolic proteins through basal autophagy is important for preventing the accumulation of abnormal proteins, which can disrupt neural function and ultimately lead to neurodegeneration.

Every eukaryotic cell has two main systems for the degradation of intracellular components: the ubiquitin–proteasome system and autophagy. Autophagy is a generic term for the degradation of cellular components in lysosomes^{1–4}. Macroautophagy (hereafter referred to as autophagy) is believed to be the main pathway among several subtypes of autophagy. During the process of autophagy, small portions of cytoplasm are sequestered by autophagosomes and then degraded on fusion with lysosomes. In contrast to the ubiquitin–proteasome system, which accounts for most of the selective intracellular protein degradation, autophagy is less selective. Autophagy induced by starvation is a mechanism for producing amino acids within cells. In yeast, autophagy-defective cells are susceptible to starvation. In comparison, mice deficient for *Atg5* and *Atg7*, which are essential for autophagosome formation⁹, suffer from severe nutrient- and energy-insufficiency soon after birth^{10,11}. Thus, adaptation to starvation is an evolutionarily conserved role of autophagy.

In addition to induced autophagy, a low level of constitutive autophagy is important for intracellular clearance under normal conditions. Mice bearing a liver-specific conditional knockout allele of *Atg7* show hepatic dysfunction and intracellular ubiquitin-positive inclusion bodies¹¹. We have also observed the accumulation of ubiquitin-positive inclusion bodies in hepatocytes and a subset of neurons in *Atg5*-knockout (*Atg5*^{−/−}) neonates (Supplementary

Fig. S1); however, conventional histological analysis revealed no significant abnormality¹⁰. These data suggest that intracellular protein quality-control by autophagy is particularly important in neural cells. Indeed, several studies have suggested that impairment of autophagy could worsen the accumulation of abnormal proteins in neurodegenerative disease models *in vitro* and *in vivo*^{5–8}. However, direct evidence demonstrating that autophagy contributes to the prevention of neurodegeneration has been lacking, in part because *Atg5*^{−/−} and *Atg7*^{−/−} mice die soon after birth^{10,11}.

To determine the role of autophagy in neural cells, we generated neural-cell-specific *Atg5*^{−/−} mice (Supplementary Fig. S2). Mice bearing an *Atg5*^{lox} allele, in which exon 3 of the *Atg5* gene is flanked by two *loxP* sequences, were crossed with a transgenic line expressing Cre recombinase under the control of the nestin promoter (*nestin-Cre*)¹². In these mice, Cre recombinase is expressed in neural precursor cells after embryonic day (E)10.5, causing deletion of the *loxP*-flanked exon 3 (Supplementary Fig. S3). Recombination was successful in over 90% of all brain cells from *Atg5*^{lox/lox}; *nestin-Cre* mice. The expression of *Atg5* (detected as an *Atg12*–*Atg5* conjugate¹³) and the *Atg5*-dependent conversion of microtubule-associated protein 1 light chain 3 (LC3)-I to LC3-II (LC3–phosphatidylethanolamine (LC3–PE) conjugate)^{13,14} were almost completely suppressed in the brains of *Atg5*^{lox/lox}; *nestin-Cre* mice after E15.5 (Fig. 1a and Supplementary Fig. S3). These data suggest that autophagosome formation is impaired in the brains of these mutant mice.

Atg5^{lox/lox}; *nestin-Cre* mice were born normally and survived neonatal starvation. They did not show the suckling defects observed in *Atg5*^{−/−} and *Atg7*^{−/−} neonates^{10,11}, suggesting that an undetectable, but sufficient, level of *Atg5* remains in the neurons controlling the suckling response at this stage, or that non-neural cells may mediate the suckling deficit in the non-conditional mutants. However, the *Atg5*^{lox/lox}; *nestin-Cre* mice showed growth retardation: their mean body weight was about 1.5-times lower than that of control (*Atg5*^{lox/+}; *nestin-Cre*) mice (Fig. 1b). *Atg5*^{lox/lox}; *nestin-Cre* mice developed progressive motor and behavioural deficits after three weeks of age, and footprint analysis revealed an ataxic walking pattern (Fig. 1c). Mean stride lengths corrected for paw base widths were significantly decreased compared with control (*Atg5*^{lox/+}; *nestin-Cre*) mice. The *Atg5*^{lox/lox}; *nestin-Cre* mice showed limb-clasping reflexes when they were suspended by their tails, whereas control mice extended their limbs (Fig. 1d). This abnormal reflex is often observed in mouse models of neurodegenerative disease^{15,16}.

¹Department of Bioregulation and Metabolism, Tokyo Metropolitan Institute of Medical Science, Tokyo 113-8613, Japan. ²Mouse Genome Technology Laboratory, Mitsubishi Kagaku Institute of Life Sciences, Tokyo 194-8511, Japan. ³Department of Basic Biology, School of Life Science, the Graduate University for Advanced Studies, Okazaki 444-8585, Japan. ⁴Department of Cell Biology, National Institute for Basic Biology, Okazaki 444-8585, Japan. ⁵Department of Bio-Science, Nagahama Institute of Bio-Science and Technology, Nagahama 526-0829, Japan. ⁶Brain Research Institute, Niigata University, Niigata 951-8510, Japan. ⁷Department of Pathology, Tsurumi University School of Dental Medicine, Yokohama 230-8501, Japan. ⁸Department of Physiology, Keio University School of Medicine, Tokyo 160-8582, Japan. ⁹SORST and ¹⁰PRESTO, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan.

Rotarod (Fig. 1e) and wire-hanging (data not shown) tasks also showed severely impaired motor coordination, balance and grip strength in *Atg5^{flox/flox}; nestin-Cre* mice. Finally, tremor was apparent in 12-week-old mice. Some of the *Atg5^{flox/flox}; nestin-Cre* mice died after three weeks of age. Neither *Atg5^{flox/flox}* mice (Cre-negative) nor *Atg5^{flox/+}; nestin-Cre* mice showed any abnormal phenotype.

The gross anatomy of the brain of the mutant mice was normal. However, histological examination revealed degenerative changes in the neurons of *Atg5^{flox/flox}; nestin-Cre* mice. These alterations were most prominent in cerebellar Purkinje cells. Haematoxylin and eosin (H&E) staining (arrowheads in Fig. 2a) and immunohistochemical staining with an antibody directed against calbindin (a selective marker for Purkinje cells, left panels in Fig. 2b), demonstrated partial loss of these neurons (Fig. 2c). The remaining Purkinje cells showed eccentrically located nuclei, with infolding of the nuclear membrane. We also found a number of eosinophilic spheroids in H&E-stained sections in the cerebellar nuclei of these mutant mice (arrows in Fig. 2a), which probably correspond to the calbindin-positive spheroids in the same region (Fig. 2b, right panels). These structures suggest massive swelling of Purkinje cell axons that project to the cerebellar nuclei^{17,18}. In addition, TUNEL-positive cells were detected in the adjacent granular layer, suggesting apoptosis of granular cells in *Atg5^{flox/flox}; nestin-Cre* mice (Fig. 2d). Consistent with previous

observations, the survival of granular cells largely depends on their synaptic connectivity with Purkinje cells¹⁹. Axonal swelling was observed in other regions of the *Atg5^{flox/flox}; nestin-Cre* brain, including the cerebral cortex, the nucleus gracilis (Fig. 2a), the posterior thalamic nucleus, hippocampus, inferior colliculus, trigeminal nucleus, parabrachial nucleus, anterior thalamic nucleus, caudal pons and reticular nucleus (data not shown). Partial loss of pyramidal cells also was observed in the cerebral cortex (data not shown). Together, these data suggest that *Atg5^{flox/flox}; nestin-Cre* mice suffer from neurodegeneration.

We next examined protein aggregation in the brain using an antibody against ubiquitin, a marker of misfolded proteins. Large, ubiquitin-positive inclusion bodies accumulated in the cytoplasm of large neurons in the thalamus, pons, medulla, dorsal root ganglion (DRG) (Fig. 3a) and midbrain (data not shown) of *Atg5^{flox/flox}; nestin-Cre* mice. Neurons in the cerebral cortex, hippocampus (especially in the CA3 and CA4 regions) (Fig. 3a), striatum and olfactory bulb (data not shown) were also positive for these inclusion

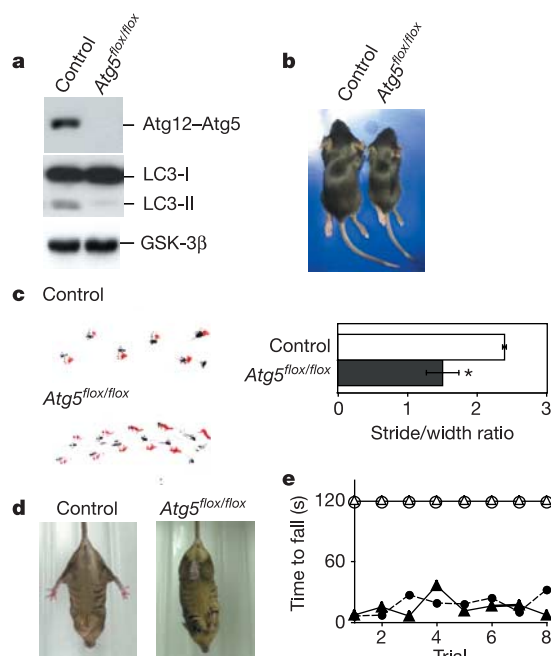


Figure 1 | Behavioural abnormalities in mice lacking *Atg5* in the nervous system. **a**, Immunoblot analysis of *Atg5* and LC3. Brain homogenates were prepared from six-week-old control (*Atg5^{flox/+}; nestin-Cre*) and *Atg5^{flox/flox}; nestin-Cre* (*Atg5^{flox/flox}*) mice. Immunoblot analysis was performed using antibodies against *Atg5* and LC3. GSK-3β was used as a loading control. The positions of the *Atg12-Atg5* conjugate, LC3-I and LC3-II (LC3-PE conjugate) are indicated. *Atg5* monomer was not detected in either lane (data not shown). **b**, A representative male control (*Atg5^{flox/+}; nestin-Cre*) and an *Atg5^{flox/flox}; nestin-Cre* (*Atg5^{flox/flox}*) littermate at three weeks of age. **c**, Left, paw placement records of eight-week-old mice. Right, stride lengths corrected for paw base widths (stride/width ratio) of *Atg5^{flox/flox}; nestin-Cre* and control (*Atg5^{flox/+}; nestin-Cre*) littermate mice. Values represent means $\pm$ s.d. of four mice. Asterisk, $P < 0.01$ (Student's *t*-test). **d**, Abnormal limb-clasping of an *Atg5^{flox/flox}; nestin-Cre* mouse compared with a control mouse (*Atg5^{flox/+}; nestin-Cre*) when suspended by its tail. **e**, Rotarod testing of *Atg5^{flox/+}; nestin-Cre* (open symbols) and *Atg5^{flox/flox}; nestin-Cre* (closed symbols) mice. One male and one female mouse were analysed for each genotype. The time until drop from the rod (rotating at 20 r.p.m.) is shown.

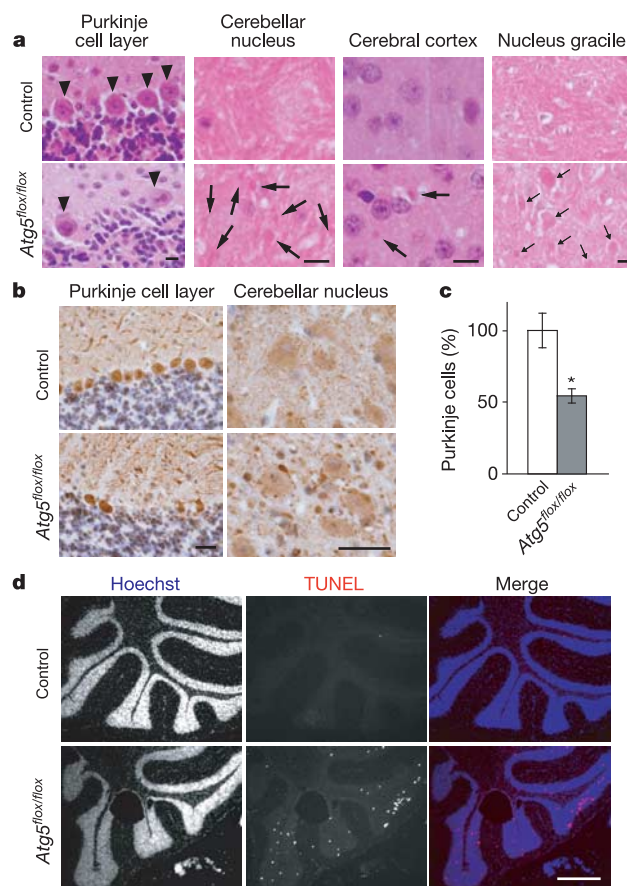


Figure 2 | Neuronal degeneration in *Atg5^{flox/flox}; nestin-Cre* mice. **a**, H&E-stained sections of the cerebral cortex, the gracile nucleus and cerebellum from control (*Atg5^{flox/+}; nestin-Cre*) and *Atg5^{flox/flox}; nestin-Cre* (*Atg5^{flox/flox}*) mice at three months of age. Purkinje cells are indicated with arrowheads. Arrows indicate eosinophilic spheroids, which represent axon swelling. Scale bar, 10 μ m. **b**, Immunohistochemistry using an anti-calbindin antibody on cerebellum sections from control (*Atg5^{flox/+}; nestin-Cre*) and *Atg5^{flox/flox}; nestin-Cre* mice at six weeks of age. Scale bar, 25 μ m. **c**, Loss of Purkinje cells in *Atg5^{flox/flox}; nestin-Cre* mice. Purkinje cells were counted in comparable areas for each mouse, and three fields were counted in each area for each mouse. Data are normalized against values from control mice (*Atg5^{flox/+}; nestin-Cre*). Values represent mean $\pm$ s.d. of three mice. Asterisk, $P < 0.01$ (*t*-test). **d**, Apoptotic death of granular cells. Cerebellum sections from control (*Atg5^{flox/+}; nestin-Cre*) and *Atg5^{flox/flox}; nestin-Cre* mice at six weeks of age were subjected to TUNEL staining. Nuclei were stained with Hoechst 33258. Scale bar, 500 μ m.

bodies. Such inclusion bodies were not observed in the brains of control ($Atg5^{flox/+}$; nestin-*Cre*) mice. We observed ubiquitin-positive inclusion bodies exclusively in cells positive for the neural cell marker NeuN, suggesting that inclusion bodies were generated only in neurons and not in glial cells (Fig. 3b). Notably, although there was extensive loss of Purkinje cells, these neurons had very few inclusion bodies in their cell bodies (Fig. 3a). Numerous ubiquitin-positive dots were observed in the cerebellar nuclei (Fig. 3a), but most of them did not colocalize with the calbindin dots observed in Fig. 2b (data not shown).

The accumulation of inclusion bodies was time-dependent, and the distribution of inclusion-body-positive cells was more limited in $Atg5^{-/-}$ and $Atg5^{flox/flox}$; nestin-*Cre* newborns compared to adult mice. In $Atg5^{-/-}$ neonates, inclusion bodies were observed in the pons, DRG, spinal cord (ventral horn) (Supplementary Fig. S1), hypothalamus, midbrain and trigeminal ganglia (data not shown), but not in the cerebral cortex (Supplementary Fig. S1). A similar pattern was observed in the brain and DRG of $Atg5^{flox/flox}$; nestin-*Cre* neonates (Supplementary Fig. S4 and data not shown). Immunoelectron microscopy of DRG neurons isolated from $Atg5^{-/-}$ neonates

demonstrated the specific association of ubiquitin with amorphous, occasionally filamentous, structures (Fig. 3c, left), as well as with more compact structures surrounded by filamentous materials (Fig. 3c, right; see Supplementary Fig. S5 for larger images). Outside the brain, $Atg5^{-/-}$ neonates showed inclusion body formation in the liver, the anterior lobe of pituitary gland (Supplementary Fig. S1) and the adrenal gland (data not shown).

Histological examination of $Atg5^{-/-}$ and $Atg5^{flox/flox}$; nestin-*Cre* mice suggested that, in addition to the presence of inclusion bodies, the intensity of diffuse cytoplasmic ubiquitin staining was higher compared with wild-type mice. We thus analysed the time course of accumulation of diffuse ubiquitinated proteins and inclusions in DRG neurons (the neonatal tissue in which inclusion body formation was most striking). DRG neurons from $Atg5^{-/-}$ embryos at E13.5 had no apparent abnormality. However, at E15.5, some neurons had accumulated diffuse, cytosolic ubiquitinated proteins with infrequent inclusions (Fig. 4a). Later, in newborn (postnatal day (P)0) $Atg5^{-/-}$ mice, multiple ubiquitin-positive inclusion bodies were present in DRG neurons. Thus, the accumulation of diffuse abnormal proteins seems to be the primary cellular phenotype of $Atg5^{flox/flox}$; nestin-*Cre* neurons. We obtained similar results using a biochemical method with whole brains. Polyubiquitinated proteins that accumulated in $Atg5^{flox/flox}$; nestin-*Cre* brains were primarily Triton-soluble in six-week-old mice (Fig. 4b). In contrast, in 14-week-old mice, Triton-insoluble polyubiquitinated proteins were also abundant, suggesting that inclusion body formation is a later event.

We confirmed these observations using hepatocytes, another cell type that showed extensive inclusion body accumulation under autophagy-defective conditions¹¹ (Supplementary Fig. S1). We crossed $Atg5^{flox/flox}$ mice with CAG-*Cre* mice that express Cre recombinase ubiquitously (see Methods)²⁰. As recombination efficiency was not high when crossed with our $Atg5^{flox/flox}$ mice, the resulting mice ($Atg5^{flox/flox}$; CAG-*Cre*) were mosaic for the mutant allele and viable. In the livers of these mice, the *Atg5* gene was deleted in only about 30% of hepatocytes (Supplementary Fig. S6), which allowed us to compare directly the immunoreactivity of knockout cells with that of wild-type cells in the same specimen. Anti-ubiquitin staining of the liver in four-month-old $Atg5^{flox/flox}$; CAG-*Cre* mice showed that about 30% of cells had very high levels of diffuse cytoplasmic signals (in addition to inclusion bodies), compared with surrounding cells that were probably wild type (Fig. 4c). The results clearly demonstrate that cytosolic ubiquitinated proteins also accumulate in $Atg5^{-/-}$ hepatocytes.

We then determined the time course of accumulation of ubiquitinated proteins using Mx1-*Cre* transgenic mice²¹. In this system, Cre recombinase is expressed under the control of an interferon-responsive promoter that can be activated by application of polyinosinic acid-polycytidylic acid (pIpC), an interferon-inducible, synthetic double-stranded RNA. The *Atg5* gene in $Atg5^{flox/flox}$; Mx1-*Cre* liver was deleted by intraperitoneal injection of pIpC. Soon after injection, most targeted cells showed only diffuse ubiquitin staining, with no inclusion bodies (Fig. 4d, e). In contrast, large, ubiquitin-positive inclusion bodies were present in almost all targeted hepatocytes 16 days after pIpC injection (arrowheads in Fig. 4d). Cells with only inclusion bodies but not diffuse cytoplasmic staining were observed very rarely. Taken together, our data demonstrate that loss of autophagy first leads to accumulation of diffuse abnormal proteins, followed by generation of inclusion bodies.

We have shown that the inhibition of autophagy in neural cells causes neurodegeneration and symptoms of neurological pathology. As the mouse model we used does not express any disease-associated mutant proteins, the phenotype of these mutant mice indicates that autophagy mediates essential and continuous turnover of intracellular proteins. This system is particularly important for neurons, in which deregulation of this degradation process can induce cell dysfunction. The role of autophagy could be even more critical if

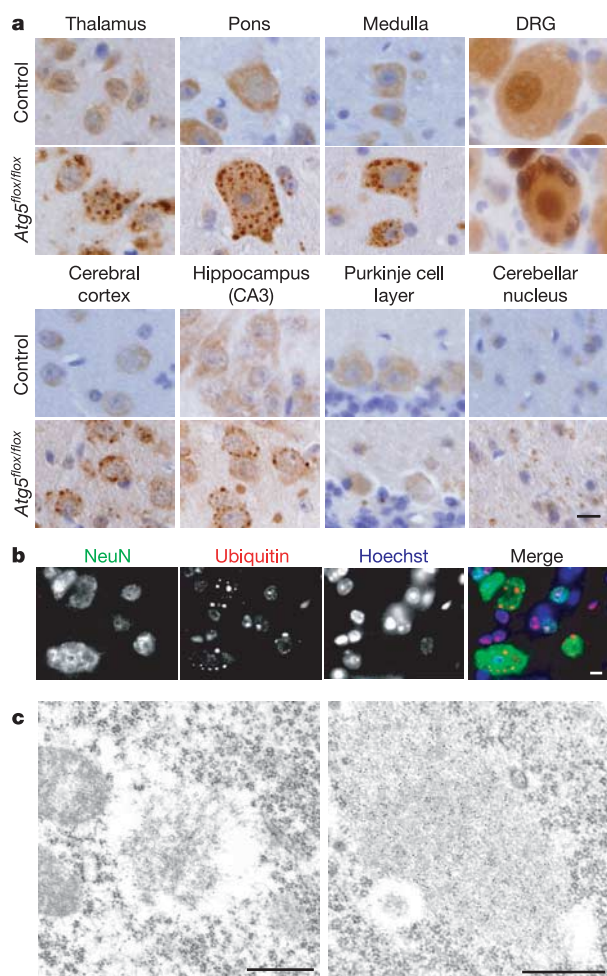


Figure 3 | Ubiquitin-positive inclusions in *Atg5*-deficient neurons. **a**, Immunohistochemistry of brain sections from control ($Atg5^{flox/+}$; nestin-*Cre*) and $Atg5^{flox/flox}$; nestin-*Cre* ($Atg5^{flox/flox}$) mice at six weeks of age, stained with an anti-ubiquitin antibody (1B3). Scale bar, 10 μ m. **b**, Ubiquitin-positive inclusions in neurons. Sections of medulla from $Atg5^{flox/flox}$; nestin-*Cre* mice at six weeks of age were stained with an antibody against NeuN (a neuron-specific nuclear protein) and ubiquitin. Nuclei were stained with Hoechst 33258. Scale bar, 10 μ m. **c**, Immunoelectron microscopy of ubiquitin-positive inclusion bodies. DRG neurons isolated from $Atg5^{-/-}$ neonates were analysed by immunoelectron microscopy using an anti-ubiquitin antibody. Scale bars, 500 nm.

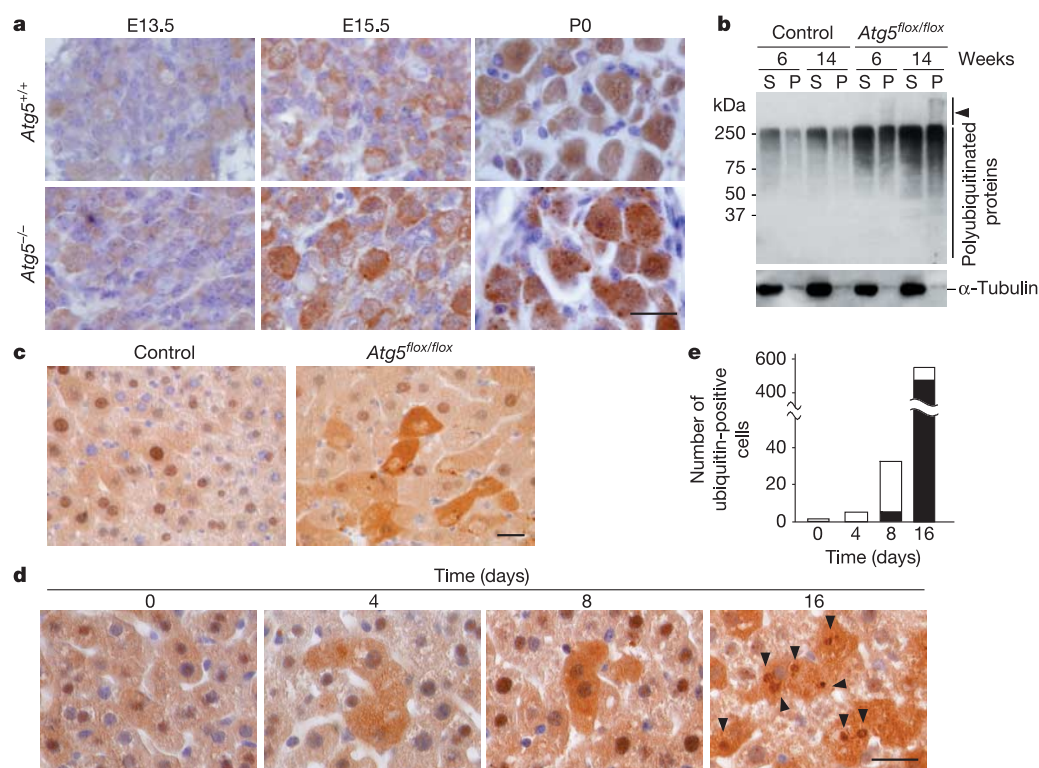


Figure 4 | Accumulation of diffuse ubiquitinated proteins in autophagy-defective cells. **a**, Immunohistochemistry of DRG neurons from *Atg5^{+/+}* and *Atg5^{-/-}* mice at different developmental stages (embryonic day (E) 13.5, E15.5 or postnatal day (P) 0), stained with an antibody directed against ubiquitin. Scale bar, 25 μ m. **b**, Accumulation of Triton-X-100-soluble polyubiquitinated proteins in the brains of *Atg5^{flox/flox}*; nestin-Cre mice. Brain homogenate prepared at the indicated times from control (*Atg5^{flox/+}*; nestin-Cre) and *Atg5^{flox/flox}*; nestin-Cre (*Atg5^{flox/flox}*) mice were separated into Triton-X-100-soluble (S) and -insoluble (P) fractions and analysed by immunoblotting using anti-ubiquitin antibodies. Arrowhead indicates the

stacking gel. **c**, Immunohistochemistry of liver sections from control (*Atg5^{flox/+}*; CAG-Cre) and *Atg5^{flox/flox}*; CAG-Cre (*Atg5^{flox/flox}*) mice at four months of age, using an anti-ubiquitin antibody. Scale bar, 25 μ m.

d, Immunohistochemistry of liver sections from six-week-old *Atg5^{flox/flox}*; Mx1-Cre mice at the indicated time points after plpC injection, using anti-ubiquitin antibodies. Arrowheads indicate ubiquitin-positive inclusion bodies. Scale bar, 25 μ m. **e**, Five thousand hepatocytes were randomly selected, and the number of cells with diffuse cytosolic ubiquitin signals, with (black) or without (white) inclusion bodies, were counted.

any aggregation-prone mutant proteins were expressed. Although autophagy is generally thought to be a non-selective process, several studies have suggested that autophagosomes can specifically engulf inclusion bodies^{8,22}. In our system, however, inclusion bodies appeared in later phases of autophagy deficiency, suggesting that the primary role of autophagy under normal conditions is the turnover of diffuse cytosolic proteins, not direct elimination of inclusion bodies. As the population of ubiquitinated proteins in *Atg5^{flox/flox}*; nestin-Cre brains was similar to that in wild-type mouse brains, we suggest that cytoplasmic proteins that are usually ubiquitinated, rather than specific proteins, accumulate in larger amounts in the absence of autophagy (Supplementary Fig. S7). Downregulation of protein turnover could cause the accumulation of abnormal proteins, which then could promote aggregate formation (Supplementary Fig. S8).

The critical role of autophagy in the basal turnover of diffuse cytosolic proteins in neural cells should be emphasized, because it has been suggested that large inclusion bodies themselves might not be pathogenic, but that mutant proteins present diffusely in the cytosol could be the primary source of toxicity^{23–27}. However, we do not rule out the possibility that autophagosomes can selectively recognize abnormal soluble proteins or microaggregates on the membrane surface. It was recently reported that the polyubiquitin-binding protein p62/SQSTM1 might mediate the specific recognition of protein aggregates by autophagosomes²⁸. This pathway might also be involved in the degradation of diffuse ubiquitinated proteins by autophagy.

METHODS

Generation of tissue-specific *Atg5*-deficient mice. An approximately 1-kb *XbaI*–*SpeI* mouse genomic fragment containing putative exon 3 of the *Atg5* gene was flanked by two *loxP* sites containing the neomycin-resistant (*neo^r*) cassette from pMC1-Neo (Stratagene). The diphtheria toxin A (*DT-A*) gene was inserted downstream of the short arm, for negative selection against random integration of the vector (Supplementary Fig. S2). Targeted CCE embryonic stem cells of 129/SvEv mouse origin were injected into C57BL/6 blastocysts, and chimaeric mice were crossed with C57BL/6 mice to obtain *Atg5^{flox/+}* mice. We used the following primers to detect wild-type *Atg5* and *Atg5^{flox}* alleles: A (exon3-1), 5'-GAATATGAAGGCACACCCCTGAAATG-3'; B (short2), 5'-GTACTGCATAATGGTTTAACTCTTGC-3'; C (check2), 5'-ACAACGTCGAGCACAGCTGCGCAAGG-3'; D (5L2), 5'-CAGGGAATGGTGTCTCCAC-3'; E (cre1), 5'-AGGTTCTGTTCACTCATGGA-3'; F (cre2), 5'-TCGACCAGTTT AGTTACCC-3'.

Southern blot analysis was performed using the probe shown in Supplementary Fig. S2 after digestion of genomic DNA with *EcoRV* and *KpnI*, as described previously¹³. Nestin-Cre transgenic mice expressing Cre recombinase under the control of the mouse nestin gene promoter and second intronic neural enhancer (a gift from S. Noguchi) have been described previously²⁹. CAG-Cre transgenic mice expressing Cre recombinase under the control of the CAG (CMV enhancer and chicken β -actin) promoter have been described previously²⁰. Mx1-Cre transgenic mice were obtained from the Jackson Laboratory²¹. Progeny containing the *Atg5^{flox}* allele were bred with these Cre transgenic mice to generate *Atg5^{flox/flox}*; nestin-Cre, *Atg5^{flox/flox}*; CAG-Cre and *Atg5^{flox/flox}*; Mx1-Cre mice. *Atg5^{-/-}* mice have been described previously¹⁰. All animal experiments were approved by the institutional committee of the Tokyo Metropolitan Institute of Medical Science.

Antibodies. A monoclonal antibody against ubiquitin (1B3) was purchased from

MBL and used for immunohistochemistry. Rabbit anti-ubiquitin polyclonal antibody (DakoCytomation) was used for immunoelectron microscopy. Anti-polyubiquitin monoclonal antibody (FK2, Nippon Bio-Test Laboratories) was used in immunoblot analyses. The following antibodies were also used: anti-NeuN monoclonal antibody (Chemicon), rabbit anti-calbindin polyclonal antibody (Chemicon), Alexa Fluor 488- and 660-conjugated goat anti-rabbit IgG (H + L) antibodies (Molecular Probes), monoclonal anti-glycogen synthase kinase-3 β antibody (BD Biosciences), monoclonal anti- α -tubulin antibody (DM1A, Sigma-Aldrich), and antibodies against Atg5 (SO4)¹³ and LC3¹⁴.

Behavioural analysis. Mice were placed on a rod rotating at 20 r.p.m., and the time taken for them to fall from the rod was measured. If a mouse stayed on the rod until the end of the 2-min trial, a time of 120 s was recorded.

Immunohistochemical analysis. Mice were transcardially perfused with 4% paraformaldehyde in phosphate buffer (pH 7.4). Tissues were post-fixed in the same fixative overnight and embedded in paraffin. Sections were stained using Meyer's H&E. For immunohistochemical analysis, all tissue sections were subjected to antigen retrieval using the microwave method (in 0.01 M citrate buffer for 10 min). After blocking, sections were incubated with primary antibodies for 1 h, followed by 30 min incubation with fluorescently labelled or biotinylated secondary antibodies that were detected using Histomouse-plus kits (Zymed Laboratories) and the Liquid DAB substrate chromogen system (DakoCytomation). Apoptotic cells were detected by TUNEL assay using an *in situ* cell death detection kit (Roche Diagnostic).

Immunoelectron microscopy. For immunoelectron microscopy, the post-embedding immuno-gold method was used to label tissue sections embedded with LR white resin (London Resin Co.) as previously described³⁰.

Preparation of detergent-soluble and -insoluble fractions. Mouse brains were homogenized in five volumes of ice-cold 0.25 M sucrose buffer (50 mM Tris-HCl pH 7.4, 1 mM EDTA) with protease inhibitors. Homogenates were centrifuged at 500g for 10 min at 4 °C, and the resulting supernatants were lysed with an equal volume of cold sucrose buffer containing 1% Triton X-100. Lysates were subjected to centrifugation at 13,000g for 15 min at 4 °C to separate supernatants (fractions soluble in 0.5% Triton-X-100) and pellets. Pellets were resuspended in 1% SDS in PBS (Triton-X-100-insoluble fractions).

Received 6 February; accepted 20 March 2006.

Published online 19 April 2006.

- Cuervo, A. M. Autophagy: in sickness and in health. *Trends Cell Biol.* **14**, 70–77 (2004).
- Levine, B. & Klionsky, D. J. Development by self-digestion: molecular mechanisms and biological functions of autophagy. *Dev. Cell* **6**, 463–477 (2004).
- Klionsky, D. J. The molecular machinery of autophagy: unanswered questions. *J. Cell Sci.* **118**, 7–18 (2005).
- Mizushima, N. The pleiotropic role of autophagy: from protein metabolism to bactericide. *Cell Death Differ.* **12**, 1535–1541 (2005).
- Ravikumar, B., Duden, R. & Rubinsztein, D. C. Aggregate-prone proteins with polyglutamine and polyalanine expansions are degraded by autophagy. *Hum. Mol. Genet.* **11**, 1107–1117 (2002).
- Fortun, J., Dunn, W. A. Jr, Joy, S., Li, J. & Notterpek, L. Emerging role for autophagy in the removal of aggregates in Schwann cells. *J. Neurosci.* **23**, 10672–10680 (2003).
- Ravikumar, B. *et al.* Inhibition of mTOR induces autophagy and reduces toxicity of polyglutamine expansions in fly and mouse models of Huntington disease. *Nature Genet.* **36**, 585–595 (2004).
- Iwata, A. *et al.* Increased susceptibility of cytoplasmic over nuclear polyglutamine aggregates to autophagic degradation. *Proc. Natl Acad. Sci. USA* **102**, 13135–13140 (2005).
- Mizushima, N., Ohsumi, Y. & Yoshimori, T. Autophagosome formation in mammalian cells. *Cell Struct. Funct.* **27**, 421–429 (2002).
- Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432**, 1032–1036 (2004).
- Komatsu, M. *et al.* Impairment of starvation-induced and constitutive autophagy in Atg7-deficient mice. *J. Cell Biol.* **169**, 425–434 (2005).
- Betz, U. A., Voshchenrich, C. A., Rajewsky, K. & Muller, W. Bypass of lethality with mosaic mice generated by Cre-loxP-mediated recombination. *Curr. Biol.* **6**, 1307–1316 (1996).
- Mizushima, N. *et al.* Dissection of autophagosome formation using Apg5-deficient mouse embryonic stem cells. *J. Cell Biol.* **152**, 657–667 (2001).
- Kabeya, Y. *et al.* LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* **19**, 5720–5728 (2000).
- Cote, F., Collard, J. F. & Julien, J. P. Progressive neuronopathy in transgenic mice expressing the human neurofilament heavy gene: a mouse model of amyotrophic lateral sclerosis. *Cell* **73**, 35–46 (1993).
- Mangiarini, L. *et al.* Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* **87**, 493–506 (1996).
- Kikuchi, T., Mukoyama, M., Yamazaki, K. & Moriya, H. Axonal degeneration of ascending sensory neurons in gracile axonal dystrophy mutant mouse. *Acta Neuropathol. (Berl.)* **80**, 145–151 (1990).
- Sotelo, C. Axonal abnormalities in cerebellar Purkinje cells of the 'hyperspiny Purkinje cell' mutant mouse. *J. Neurocytol.* **19**, 737–755 (1990).
- Lossi, L., Mioletti, S. & Merighi, A. Synapse-independent and synapse-dependent apoptosis of cerebellar granule cells in postnatal rabbits occur at two subsequent but partly overlapping developmental stages. *Neuroscience* **112**, 509–523 (2002).
- Sakai, K. & Miyazaki, J. A transgenic mouse line that retains Cre recombinase activity in mature oocytes irrespective of the cre transgene transmission. *Biochem. Biophys. Res. Commun.* **237**, 318–324 (1997).
- Kuhn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269**, 1427–1429 (1995).
- Kipito, R. R. Aggregates, inclusion bodies and protein aggregation. *Trends Cell Biol.* **10**, 524–530 (2000).
- Saudou, F., Finkbeiner, S., Devys, D. & Greenberg, M. E. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell* **95**, 55–66 (1998).
- Kuemmerle, S. *et al.* Huntington aggregates may not predict neuronal death in Huntington's disease. *Ann. Neurol.* **46**, 842–849 (1999).
- Taylor, J. P. *et al.* Aggregates protect cells by enhancing the degradation of toxic polyglutamine-containing protein. *Hum. Mol. Genet.* **12**, 749–757 (2003).
- Arasate, M., Mitra, S., Schweitzer, E. S., Segal, M. R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* **431**, 805–810 (2004).
- Tanaka, M. *et al.* Aggregates formed by α -synuclein and synphilin-1 are cytoprotective. *J. Biol. Chem.* **279**, 4625–4631 (2004).
- Bjorkoy, G. *et al.* p62/SQSTM1 forms protein aggregates degraded by autophagy and has a protective effect on huntingtin-induced cell death. *J. Cell Biol.* **171**, 603–614 (2005).
- Mori, H. *et al.* Socs3 deficiency in the brain elevates leptin sensitivity and confers resistance to diet-induced obesity. *Nature Med.* **10**, 739–743 (2004).
- Yamamoto, A. *et al.* Stacks of flattened smooth endoplasmic reticulum highly enriched in inositol 1,4,5-trisphosphate (InsP₃) receptor in mouse cerebellar Purkinje cells. *Cell Struct. Funct.* **16**, 419–432 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements We thank H. Neko, M. Miwa and Y. Kabeya for technical assistance. We also thank J. Miyazaki for the donation of CAG-Cre transgenic mice, T. Yoshimori for the anti-LC3 antibody, E. Yamada for histological examination, M. Yuzaki for the rotarod analysis, and A. Kuma for discussion. We thank Z. Yue for critical reading of the manuscript. This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The authors thank the Yamada Science Foundation and the Cell Science Research Foundation for their financial support.

Author Contributions T.H. performed most of the experiments to characterize the neuron-specific knockout mice. M.M. analysed Atg5^{-/-} mice. K.N., Y.N., R.S.-M. and M.Y. generated Atg5^{lox} chimaeric mice. A.Y. performed electron microscopy. K.M. and I.S. performed histological analysis. H.O. provided nestin-Cre mice and participated in manuscript preparation. N.M. conceived the experiments and generated the targeting vector. T.H. and N.M. wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.M. (nmizu@rinshoken.or.jp).

LETTERS

Interleukin-2 signals during priming are required for secondary expansion of CD8⁺ memory T cells

Matthew A. Williams¹, Aaron J. Tynnik¹ & Michael J. Bevan¹

Although interleukin-2 (IL-2) was initially characterized as the primary T-cell growth factor following *in vitro* activation¹, less is known about its role in shaping T-cell responses to acute infections *in vivo*. The use of IL-2- or IL-2-receptor-deficient mice is problematic owing to their early development of autoimmunity^{2–5}, attributable to the central role of IL-2 in the generation, maintenance and function of CD4⁺CD25⁺ regulatory T cells^{6–9}. To bypass these inherent difficulties, we have studied the effect of IL-2 on T-cell responses to acute infections by adopting a mixed chimaera strategy in which T cells lacking the high-affinity IL-2 receptor could be studied in an otherwise healthy mouse containing a full complement of regulatory T cells. Here we show that although IL-2 signalling to pathogen-specific CD8⁺ T cells affects the number of developing effector and memory cells very little, it is required for the generation of robust secondary responses. This is not due to an altered T-cell-receptor repertoire development or selection, and does not reflect an acute requirement for IL-2 during secondary activation and expansion. Rather, we demonstrate a previously unappreciated role for IL-2 during primary infection in programming the development of CD8⁺ memory T cells capable of full secondary expansion. These results have important implications for the development of vaccination or immunotherapeutic strategies aimed at boosting memory T-cell function.

In recent years, many factors required for the generation, homeostatic turnover, and long-term survival of memory T cells have been identified, including cytokines such as IL-7 and IL-15 (refs 10–13). CD4⁺ T cells, though largely dispensable for primary CD8⁺ T cell responses to many acute infections, play a particularly crucial role in the generation of functional CD8⁺ memory T cells^{14–18}. IL-2, a product of CD4⁺ T cells, is a member of the same cytokine family as IL-7 and IL-15 by virtue of its shared usage of the common γ chain as part of its receptor and was an obvious candidate for further study. Early reports conflict on the ability of T cells to respond to infection in the setting of IL-2 deficiency^{19,20}, but these studies remain difficult to interpret, because IL-2- and IL-2R-deficient mice develop lymphoproliferation and autoimmunity at an early age^{2–5}. More recent attempts to study IL-2 in healthy hosts have found that it plays a surprisingly modest role in driving the development of effector cytotoxic T lymphocytes^{21–23}.

We aimed to study the long-term impact of IL-2 signalling on the development of CD8⁺ memory T-cell numbers and function in mice that did not suffer from ongoing autoimmune disease. To create this setting, we made mixed bone-marrow chimaeras in which irradiated C57BL/6 (B6) mice (CD45.1⁺) were reconstituted with T-cell-depleted bone marrow from B6 (Thy1.1⁺) and IL-2R α -deficient (Thy1.2⁺) donors (hereafter referred to as WT/IL-2R α ^{-/-} chimaeras), thus allowing us to track wild-type (WT) and IL-2R α -deficient T-cell responses in a healthy mouse with a full complement of regulatory T cells (Supplementary Fig. 1).

Following acute infection of WT/IL-2R α ^{-/-} chimaeras with

lymphocytic choriomeningitis virus (LCMV), both WT and IL-2R α -deficient CD8⁺ T cells generated comparable populations of effector and memory cells specific for the immunodominant epitope of the LCMV glycoprotein, GP_{33–41} (Fig. 1a). Similar results were seen for another immunodominant CD8 epitope, NP_{396–404}, and for the immunodominant CD4 epitope GP_{61–80} (data not shown). However, following a rechallenge with *Listeria monocytogenes* secreting the GP_{33–41} epitope (LM-GP), IL-2R α -deficient memory cells mounted highly impaired secondary responses (Fig. 1a). Whereas the frequencies of WT to IL-2R α -deficient memory cells were maintained at or near a ratio of 2:1, this ratio rose to 14:1 after re-challenge (Fig. 1b). Furthermore, while WT memory cells expanded nearly 40-fold after re-challenge, IL-2R α -deficient memory cells only expanded fourfold (Fig. 1c). These results were repeated when LCMV-immune chimaeras were re-challenged with a high dose of

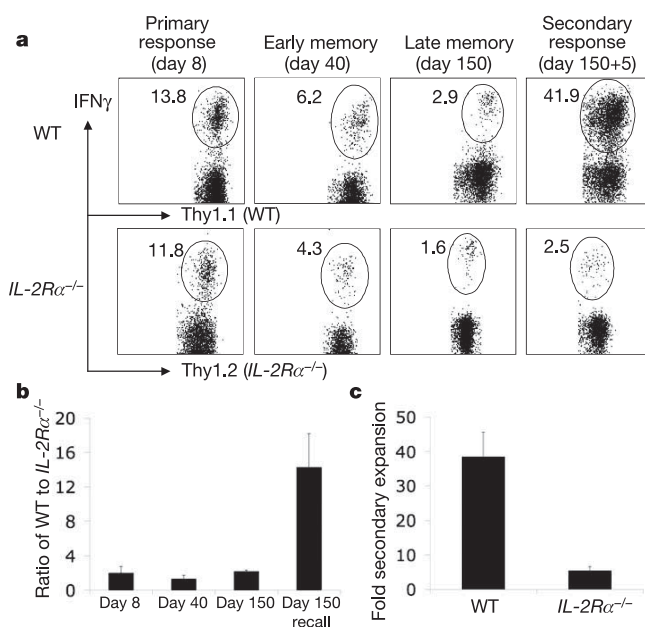


Figure 1 | IL-2R α -deficient CD8⁺ T cells generate robust primary but defective secondary responses. **a**, WT/IL-2R α ^{-/-} mixed chimaeras were infected with 2×10^5 plaque-forming units (PFU) LCMV and assessed for the frequency of interferon- γ (IFN γ)-producing cells specific for the GP_{33–41} epitope among either WT or IL-2R α -deficient CD8⁺ T cells in the spleen. In all experiments, CD45.1⁺ host cells were excluded. Mice were re-challenged 150 days post-infection with 1×10^5 colony-forming units (CFU) LM-GP, and GP_{33–41}-specific responses in the spleen were analysed 5 days later (day 150 + 5). **b**, The ratio of GP_{33–41}-specific WT to IL-2R α -deficient responders is shown at each time point. **c**, The fold expansion by 5 days after re-challenge is shown for WT and IL-2R α -deficient T cells. Error bars display s.e.m. ($n = 3–4$) and results are representative of five separate time courses.

¹Howard Hughes Medical Institute and the Department of Immunology, University of Washington, Seattle, Washington 98195, USA.

LCMV (Supplementary Fig. 2) and when chimaeras immune to a recombinant *Listeria monocytogenes* secreting ovalbumin (LM-Ova) were given a high-dose secondary challenge with LM-Ova (Supplementary Fig. 3).

To determine whether the inability of IL-2R α -deficient memory cells to respond to re-challenge was due to altered T-cell repertoire development or selection in the absence of IL-2 signalling, we measured the responses of WT and IL-2R α -deficient P14 T-cell-receptor transgenic cells (specific for the GP_{33–41} epitope of LCMV) after LCMV infection. To ensure that these cells would develop in hosts with sufficient regulatory T cells, we generated mixed chimaeras using a mixture of B6 bone marrow and either WT or IL-2R α -deficient P14 bone marrow to reconstitute lethally irradiated B6 mice. Eight weeks later, CD44^{lo} P14 T cells were harvested from these chimaeras and co-transferred into new B6 hosts, followed by LCMV infection (Supplementary Fig. 4). Because the WT P14 donors were Thy1.2⁺, the IL-2R α -deficient P14 donors were Thy1.1⁺/1.2⁺, and the recipient host was Thy1.1⁺, we were able to measure the responses of WT and IL-2R α -deficient P14 cells in the same host.

Following LCMV infection, both WT and IL-2R α -deficient P14 T cells expanded robustly and developed long-term memory populations in the spleen (Fig. 2a, b) and in the mesenteric lymph nodes and liver (Supplementary Fig. 5). Furthermore, both WT and IL-2R α -deficient P14 memory cells displayed normal homeostatic turnover, as assessed by 5-bromo-2'-deoxyuridine (BrdU) incorporation over a 7-day period (Fig. 2c). Both groups of memory cells expressed similarly high levels of CD122 and CD44 (data not shown), while the frequency of cells expressing CD62L and IL-7R α was actually higher at all memory time points in the IL-2R α -deficient population than in the WT memory population (Fig. 2d). Furthermore, following *ex vivo* re-stimulation IL-2R α -deficient memory cells rapidly upregulated CD69 and contained a consistently higher frequency of IL-2-producers than WT memory cells (Supplementary Fig. 6). The phenotype of the IL-2R α -deficient cells was consistent with a central memory-like phenotype, which has been shown to correlate with high proliferative and protective capacity²⁴. The rapid loss of CD62L^{lo} cells in the IL-2R α -deficient population may reflect a previously suggested role for IL-2 in driving the differentiation of effector memory²⁵. A recent study observed that the early accumulation of cells with a central memory phenotype depends on a high precursor frequency of antigen-specific cells²⁶. We found that CD62L^{hi} cells emerged more rapidly in the IL-2R α -deficient population even when as few as 500 WT and IL-2R α -deficient P14 cells were co-transferred so that precursor frequencies approached endogenous levels (Supplementary Fig. 7).

At memory time points, WT and IL-2R α -deficient P14 cells were sorted and transferred into new B6 hosts in equal numbers. Following a secondary LCMV challenge, the WT memory cells far outpaced the IL-2R α -deficient memory cells in their ability to expand (Fig. 3a). Whereas WT P14 T cells had only a 2:1 advantage over the IL-2R α -deficient P14 T cells at the peak of the primary effector response, at the peak of the secondary response this ratio approached 20:1 (Fig. 3b). Similar results were again obtained when the precursor frequency of transferred P14 cells was similar to endogenous precursor frequencies (Supplementary Fig. 7), showing that IL-2 plays a central role for long-term protective immunity independently of T-cell repertoire development or selection. Following co-transfer of carboxyfluorescein diacetate succinimidyl ester (CFSE)-labelled WT and IL-2R α -deficient memory P14 into new hosts and re-challenge with LCMV, both groups rapidly diluted their CFSE within three days. However, even within this short time of antigen re-exposure WT P14 cells accumulated to far greater numbers than IL-2R α -deficient P14 cells (Fig. 3c), indicating that IL-2 is not required to drive proliferation of memory cells following secondary antigen encounter but instead promotes the survival and accumulation of dividing cells.

We hypothesized that IL-2 signalling could play a role in promoting

secondary T-cell expansion during three separate phases of the immune response: during the primary response, during memory maintenance, or following re-challenge. To address this question, we took advantage of the recent observation that the anti-IL-2 antibody S4B6, previously thought to be neutralizing, actually enhances the potency of IL-2 *in vivo* for T cells bearing the intermediate affinity receptor for IL-2, IL2R $\beta\gamma$ (ref. 27). In that study, co-administration of small amounts of recombinant IL-2/anti-IL-2 immune complexes resulted in the delivery of a potent signal through IL2R $\beta\gamma$ that was independent of IL-2R α . To determine whether IL-2/anti-IL-2

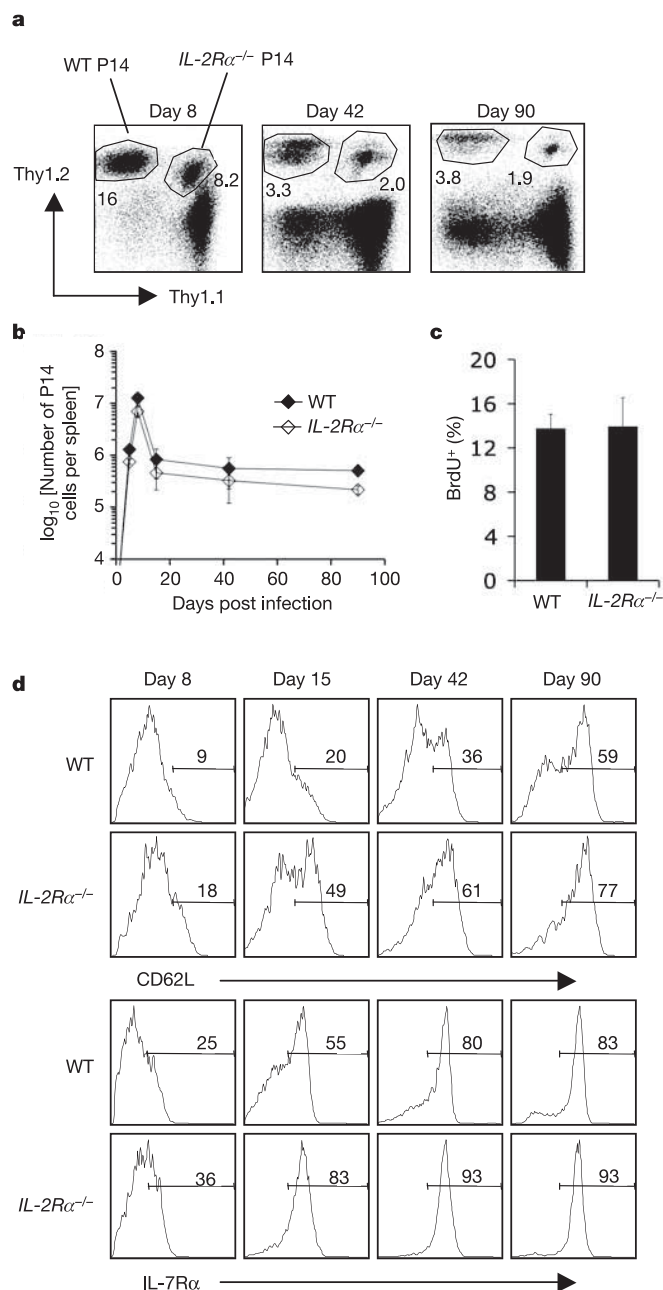


Figure 2 | IL-2R α -deficient memory cells are maintained at normal levels.

a, 1×10^4 naive WT and IL-2R α -deficient P14 cells were co-transferred into B6 hosts, infected with LCMV, and frequencies of P14 cells were measured in the spleen. **b**, The total number of WT or IL-2R α -deficient P14 cells in the spleen is shown over a 90 day time course. **c**, At day 90 post-infection, mice were fed BrdU in their drinking water for 7 days and stained for BrdU incorporation by WT or IL-2R α -deficient P14 cells. **d**, Cell surface expression of the indicated molecules by WT or IL-2R α -deficient P14 cells at days 8, 15, 42 and 90 post-infection. Results are representative of 3 separate time courses and error bars display s.e.m. ($n = 3-4$).

immune complexes could deliver functional IL-2 signals during acute infection, we injected 1×10^4 naive WT and IL-2R α -deficient P14 cells into B6 mice as described previously (Supplementary Fig. 3), infected them with LCMV and treated mice with IL-2/anti-IL-2 immune complexes or with anti-IL-2 alone. WT and IL-2R α -deficient P14 cells expanded to similar levels by day 8 post-infection in all treatment groups (data not shown). As seen previously (Fig. 2d), by day 19 post-infection, 35–40% of IL-2R α -deficient cells expressed CD62L in untreated mice, compared to 10–13% of WT P14 cells. In contrast, IL-2R α -deficient P14 cells in mice treated during the primary response with IL-2/anti-IL-2 immune complexes or with anti-IL-2 alone had levels of CD62L expression that were similar to WT P14 cells (Fig. 4a), indicating that the signalling of IL-2/anti-IL-2 immune complexes through IL2R $\beta\gamma$ can provide a functional IL-2 signal during acute infection.

Because the activity of the complexes is short-lived²⁷, we were able to design experiments in which the timing of IL-2 signals to IL-2R α -deficient T cells could be manipulated. WT/IL-2R α ^{-/-} mixed chimaeras were infected with LCMV and either left untreated,

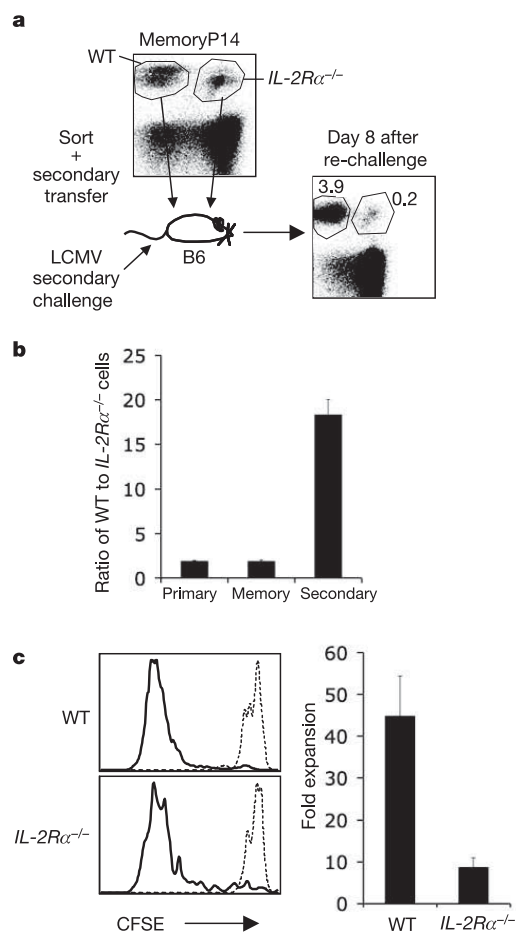


Figure 3 | IL-2R α -deficient memory cells proliferate but do not accumulate following rechallenge. **a**, At 42 days post-infection, WT and IL-2R α -deficient memory P14 cells were sorted, transferred into new B6 hosts (1×10^4 of each) and re-challenged with LCMV. Eight days after secondary challenge, we measured their frequency in the spleen. Results were consistent for four separate experiments. **b**, The ratio of WT to IL-2R α -deficient P14 cells was assessed at days 8 and 42 post-infection, as well as day 8 after re-challenge. Error bars display s.e.m. ($n = 3$ –5) and are representative of four separate experiments. **c**, At day 90 post-infection, WT and IL-2R α -deficient memory P14 cells were labelled with CFSE, transferred to a new host and re-challenged with LCMV. Flow plots indicate CFSE divisions after three days, with dashed lines representing uninfected controls, and the graph displays the fold expansion of each group. Error bars display s.e.m. ($n = 3$).

injected with anti-IL-2 alone or injected with IL-2/anti-IL-2 complexes daily for the first six days of infection. Both WT and IL-2R α -deficient CD8 T cells developed comparable effector and memory populations (data not shown). Six weeks after infection mice were re-challenged with a high dose of LM-GP, and further treatment groups received daily injections of IL-2/anti-IL-2 immune complexes or anti-IL-2 alone during the secondary response. Treatment with IL-2/anti-IL-2 immune complexes or with anti-IL-2 alone during the primary response rescued the ability of IL-2R α -deficient

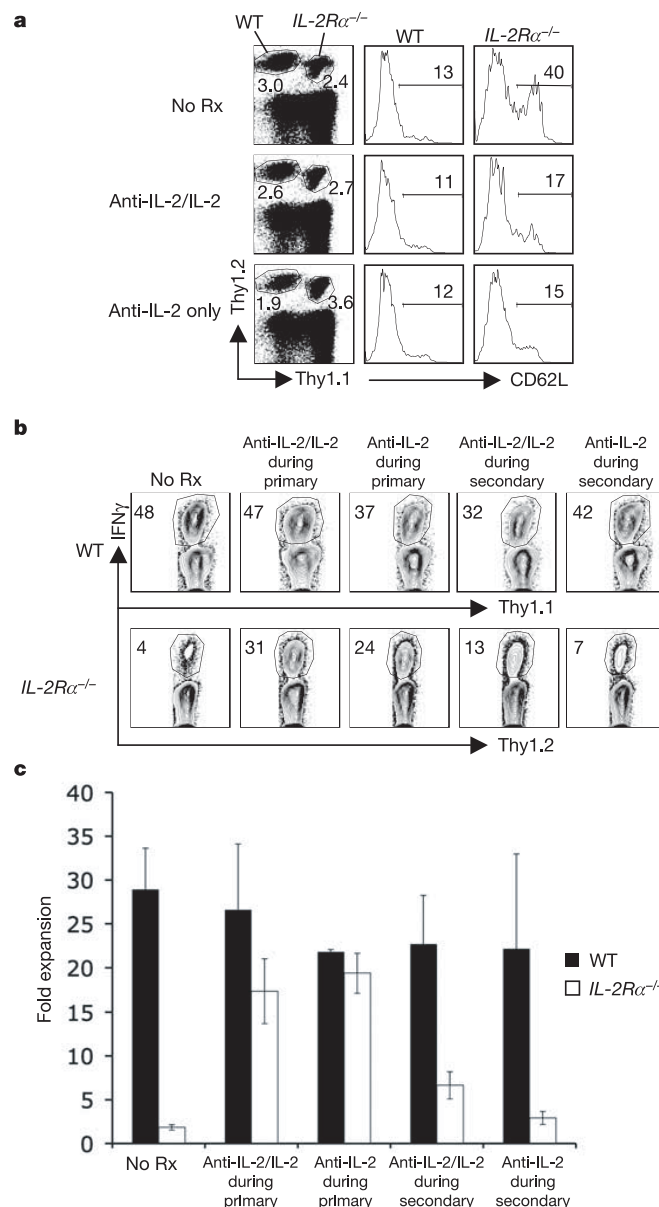


Figure 4 | IL-2 signalling during the primary response promotes secondary CD8⁺ T cell responsiveness. **a**, Following co-transfer of 1×10^4 naive WT and IL-2R α -deficient P14 cells, B6 hosts were infected with LCMV and received daily intraperitoneal injections of 50 μ g anti-IL-2 (clone S4B6) alone, daily co-injections of anti-IL-2 and 1.5 μ g recombinant mouse IL-2 on days 0–6 of the primary infection, or no treatment (No Rx). Plots display the frequency and CD62L expression of WT and IL-2R α -deficient P14 cells at day 19 post-infection. **b**, WT/IL-2R α ^{-/-} mixed chimaeras were infected with LCMV and re-challenged with LM-GP 6 weeks later. Mice were treated on either days 0–6 of the primary infection or days 0–4 of the secondary infection. Plots display the frequency of GP_{33–41}-specific WT or IL-2R α -deficient CD8⁺ T cells at day 5 post-rechallenge. **c**, The graph displays the fold expansion by 5 days post-rechallenge. Error bars display s.e.m. ($n = 3$).

CD8 memory cells to respond to re-challenge. Conversely, treatment with immune complexes during the secondary response only modestly improved the expansion of IL-2R α -deficient CD8 memory cells, while treatment with anti-IL-2 alone had no effect (Fig. 4b,c). We conclude that IL-2 signalling during the primary response programmed the formation of fully responsive CD8 memory cells capable of generating robust recall responses.

To determine whether IL-2 signals to CD8⁺ T cells were autocrine, we generated mixed chimaeras with a mix of B6 and IL-2-deficient bone marrow. In this scenario, IL-2-deficient T cells are dependent on paracrine IL-2 signalling. We observed no differences in either the primary or secondary responses of WT and IL-2-deficient T cells, demonstrating that paracrine IL-2 signalling is sufficient for the development of fully functional CD8⁺ memory cells (Supplementary Fig. 8). What cell type, then, is the source of IL-2 for promoting the development of functional CD8⁺ memory T cells? One candidate is the CD4 subset, because activated CD4⁺ T cells are the major source of IL-2 *in vivo*. Alternatively, IL-2 may come from other sources, including other activated CD8⁺ T cells, as well as activated dendritic cells²⁸.

Signals delivered during the programming of immune responses can shape the long-term fate and function of CD8 memory cells^{15,16}, and we show here that IL-2 signalling during the primary response to acute infection affects the ability of the subsequently arising CD8⁺ memory T cell population to survive and accumulate following secondary antigen exposure. Future studies should focus on the specific pathways by which IL-2 signalling during the primary response promotes enhanced survival of CD8⁺ T cells during the recall response. More experiments are also required to delineate the role of IL-2 in a broad spectrum of *in vivo* immune responses, including those directed towards other types of acute and chronic infections, tumours, and protein and DNA immunizations. Resolving these questions will be an important step in understanding the mechanisms by which fully functional CD8⁺ memory T cells are generated.

METHODS

Mice and infections. 6–8-week-old C57BL/6 (Thy1.2⁺), B6.SJL-*PtprcaPep3b*/BoyJ (CD45.1⁺) and B6.PL-*Thy1a/Cy*J (Thy1.1⁺) mice, and 4-week-old B6.129P2-*Il2tm1Hor*/J (IL-2-deficient) and B6.129S4-*Il2ratm1Dw*/J (IL-2R α -deficient) mice were purchased from Jackson Laboratories. P14 T-cell-receptor transgenic mice were bred at our facilities. Mice were infected with LCMV Armstrong, LM-Ova or LM-GP (provided by H. Shen, Univ. Pennsylvania), as described^{17,18}. Viral and bacterial stocks were prepared and propagated as described^{18,29}. Mice were injected intraperitoneally daily with 50 μ g anti-IL-2 (clone S4B6) alone or anti-IL-2 and 1.5 μ g recombinant mouse IL-2 (eBiosciences) as described²⁷.

Generation of mixed bone marrow chimaeras. Bone marrow preparations from femur and tibia were incubated with anti-CD3-biotin antibody and anti-biotin magnetic beads, followed by bead depletion on an AutoMacs (Miltenyi). 5×10^6 T-cell-depleted bone marrow cells from each of the indicated donors were injected intravenously into lethally irradiated B6 hosts (1,000 rad). Mice were infected 8–12 weeks post-transplant.

BrdU and CFSE labelling. Mice were fed 0.8 mg ml⁻¹ BrdU (Sigma) in their drinking water for 7 days and stained for BrdU incorporation according to the manufacturer's instructions (BD Pharmingen). Splenocytes were labelled in 5 μ M CFSE (Molecular Probes) for 7 min in warm RPMI.

Cell preparations and *ex vivo* restimulations. For transfer of P14 cells, CD44^{hi} cells were depleted by incubation with anti-CD44-biotin, followed by magnetic bead depletion with anti-biotin beads (Miltenyi). *Ex vivo* restimulations were performed as described³⁰ and intracellular cytokine staining was carried out according to the manufacturer's instructions (Cytofix/Cytoperm kit, BD Pharmingen).

Antibody staining and flow cytometry. Cells were stained with directly conjugated antibodies purchased from BD Pharmingen or eBiosciences and analysed on either a FACSCalibur or FACSCanto flow cytometer (Becton Dickinson). Two-way cell sorting was performed on a FACSaria (Becton Dickinson).

Received 7 March; accepted 7 April 2006.

- Smith, K. A. Interleukin-2: inception, impact, and implications. *Science* **240**, 1169–1176 (1988).
- Sadlack, B. *et al.* Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell* **75**, 253–261 (1993).

- Schorle, H., Holtschke, T., Hunig, T., Schimpl, A. & Horak, I. Development and function of T cells in mice rendered interleukin-2 deficient by gene targeting. *Nature* **352**, 621–624 (1991).
- Suzuki, H. *et al.* Deregulated T cell activation and autoimmunity in mice lacking interleukin-2 receptor β . *Science* **268**, 1472–1476 (1995).
- Willerford, D. M. *et al.* Interleukin-2 receptor α chain regulates the size and content of the peripheral lymphoid compartment. *Immunity* **3**, 521–530 (1995).
- D'Cruz, L. M. & Klein, L. Development and function of agonist-induced CD25⁺ Foxp3⁺ regulatory T cells in the absence of interleukin 2 signaling. *Nature Immunol.* **6**, 1152–1159 (2005).
- Fontenot, J. D., Rasmussen, J. P., Gavin, M. A. & Rudensky, A. Y. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nature Immunol.* **6**, 1142–1151 (2005).
- Malek, T. R. & Bayer, A. L. Tolerance, not immunity, crucially depends on IL-2. *Nature Rev. Immunol.* **4**, 665–674 (2004).
- Setoguchi, R., Hori, S., Takahashi, T. & Sakaguchi, S. Homeostatic maintenance of natural Foxp3(+) CD25(+) CD4(+) regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *J. Exp. Med.* **201**, 723–735 (2005).
- Jameson, S. C. T cell homeostasis: keeping useful T cells alive and live T cells useful. *Semin. Immunol.* **17**, 231–237 (2005).
- Kaech, S. M. *et al.* Selective expression of the interleukin 7 receptor identifies effector CD8 T cells that give rise to long-lived memory cells. *Nature Immunol.* **4**, 1191–1198 (2003).
- Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory T-cell differentiation: implications for vaccine development. *Nature Rev. Immunol.* **2**, 251–262 (2002).
- Surh, C. D. & Sprent, J. Regulation of mature T cell homeostasis. *Semin. Immunol.* **17**, 183–191 (2005).
- Bourgeois, C., Rocha, B. & Tanchot, C. A role for CD40 expression on CD8⁺ T cells in the generation of CD8⁺ T cell memory. *Science* **297**, 2060–2063 (2002).
- Janssen, E. M. *et al.* CD4⁺ T cells are required for secondary expansion and memory in CD8⁺ T lymphocytes. *Nature* **421**, 852–856 (2003).
- Shedlock, D. J. & Shen, H. Requirement for CD4 T cell help in generating functional CD8 T cell memory. *Science* **300**, 337–339 (2003).
- Sun, J. C. & Bevan, M. J. Defective CD8 T cell memory following acute infection without CD4 T cell help. *Science* **300**, 339–342 (2003).
- Sun, J. C., Williams, M. A. & Bevan, M. J. CD4⁺ T cells are required for the maintenance, not programming, of memory CD8⁺ T cells after acute infection. *Nature Immunol.* **5**, 927–933 (2004).
- Cousens, L. P., Orange, J. S. & Biron, C. A. Endogenous IL-2 contributes to T cell expansion and IFN- γ production during lymphocytic choriomeningitis virus infection. *J. Immunol.* **155**, 5690–5699 (1995).
- Kundig, T. M. *et al.* Immune responses in interleukin-2-deficient mice. *Science* **262**, 1059–1061 (1993).
- D'Souza, W. N. & Lefrançois, L. IL-2 is not required for the initiation of CD8 T cell cycling but sustains expansion. *J. Immunol.* **171**, 5727–5735 (2003).
- D'Souza, W. N., Schluns, K. S., Masopust, D. & Lefrançois, L. Essential role for IL-2 in the regulation of antiviral extralymphoid CD8 T cell responses. *J. Immunol.* **168**, 5566–5572 (2002).
- Yu, A. *et al.* Efficient induction of primary and secondary T cell-dependent immune responses *in vivo* in the absence of functional IL-2 and IL-15 receptors. *J. Immunol.* **170**, 236–242 (2003).
- Wherry, E. J. *et al.* Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nature Immunol.* **4**, 225–234 (2003).
- Manjunath, N. *et al.* Effector differentiation is not prerequisite for generation of memory cytotoxic T lymphocytes. *J. Clin. Invest.* **108**, 871–878 (2001).
- Marzo, A. L. *et al.* Initial T cell frequency dictates memory CD8⁺ T cell lineage commitment. *Nature Immunol.* **6**, 793–799 (2005).
- Boyman, O., Kovar, M., Rubinstein, M., Surh, C. D. & Sprent, J. Selective stimulation of T cell subsets with antibody-cytokine immune complexes. *Science* **311**, 1924–1927 (2006).
- Granucci, F. *et al.* Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature Immunol.* **2**, 882–888 (2001).
- Ahmed, R., Salmi, A., Butler, L. D., Chiller, J. M. & Oldstone, M. B. Selection of genetic variants of lymphocytic choriomeningitis virus in spleens of persistently infected mice. Role in suppression of cytotoxic T lymphocyte response and viral persistence. *J. Exp. Med.* **160**, 521–540 (1984).
- Murali-Krishna, K. *et al.* Counting antigen-specific CD8 T cells: a reevaluation of bystander activation during viral infection. *Immunity* **8**, 177–187 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Dere and X. Pan for technical assistance in the breeding, maintaining and typing of mouse colonies. The Howard Hughes Medical Institute and grants from the National Institutes of Health supported this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.J.B. (mbevan@u.washington.edu).

LETTERS

Signal peptide peptidase is required for dislocation from the endoplasmic reticulum

Joana Loureiro^{1†}, Brendan N. Lilley^{1†}, Eric Spooner^{2†}, Vanessa Noriega³, Domenico Tortorella³
& Hidde L. Ploegh^{1†}

Human cytomegalovirus (HCMV) prevents the display of class I major histocompatibility complex (MHC) peptide complexes at the surface of infected cells as a means of escaping immune detection¹. Two HCMV-encoded immunoevasins, US2 and US11, induce the dislocation of class I MHC heavy chains from the endoplasmic reticulum membrane and target them for proteasomal degradation in the cytosol^{2,3}. Although the outcome of the dislocation reactions catalysed is similar, US2 and US11 operate differently: Derlin-1 is a key component of the US11 but not the US2 pathway⁴. So far, proteins essential for US2-dependent dislocation have not been identified. Here we compare interacting partners of wild-type US2 with those of a dislocation-incompetent US2 mutant, and identify signal peptide peptidase (SPP) as a partner for the active form of US2. We show that a decrease in SPP levels by RNA-mediated interference inhibits heavy-chain dislocation by US2 but not by US11. Our data implicate SPP in the US2 pathway and indicate the possibility of a previously unknown function for this intramembrane-cleaving aspartic protease in dislocation from the endoplasmic reticulum.

Quality control of newly synthesized proteins in the endoplasmic reticulum (ER) involves an elaborate machinery that recognizes misfolded or misassembled proteins and targets many of them for degradation by the cytoplasmic ubiquitin–proteasome system⁵. A feature of this ER-associated protein degradation is the spatial separation between the targeting of substrates and their proteolysis, which requires substrate export from the ER to the cytoplasm by a process termed dislocation⁶.

Display of class I MHC molecules loaded with viral peptides at the surface of HCMV-infected cells is a prerequisite for their elimination by T lymphocytes. The dislocation and the ensuing proteasomal destruction of class I MHC heavy chain (HC) molecules mediated by the US2 and US11 immunoevasins allow HCMV to evade the host immune system^{2,3}. HC dislocation is unique in terms of the speed and specificity of degradation^{1–3,7}, yet the knowledge from the US11 pathway, in particular, has widened our understanding of the multiprotein complexes involved in ER dislocation more generally^{4,8–10}. In contradistinction, the cellular partners of US2 have not been identified.

US2 is a type I membrane glycoprotein with a non-cleavable signal sequence¹¹ (the replacement of which does not affect US2 function), a luminal domain that dictates a highly specific association with HC⁷, a transmembrane segment and a short cytosolic tail (residues 185–199). The US2 tail is essential for dislocation: US2₁₈₆, a cytosolic tail deletion mutant of US2, is dislocation-incompetent¹². We used an affinity purification approach⁴ (Fig. 1a) for the selective retrieval of interacting proteins required for the function of US2. The mouse H2-K^b signal sequence (K^bss) was used to target these constructs to the

ER membrane. Just like their untagged counterparts¹², both K^bss–HA–TEV-tagged US2 molecules, haemagglutinin (HA)–US2 (wild-type) and HA–US2₁₈₆, are correctly targeted to the ER membrane and associate with HC (Supplementary Fig. 1a, b). Moreover, their functional properties are unaffected by the addition of the amino-terminal tag: HA–US2 is functional in catalysing dislocation and the mutant HA–US2₁₈₆ is inactive, as assessed either by monitoring steady-state levels of HC (Fig. 1b, upper panel) or by the conversion of ER-resident glycosylated 43-kDa HC (HC + glycan) into cytosolic 40-kDa deglycosylated HC (HC–glycan) in the presence of proteasome inhibitor (Fig. 1b, lower panel). HC molecules that escape degradation progress through the secretory pathway to the cell surface (W6/32 immunoprecipitates, Supplementary Fig. 1b).

Digitonin lysates from cells expressing HA–US2 or HA–US2₁₈₆ were subjected to immunoprecipitation with anti-HA antibody, and

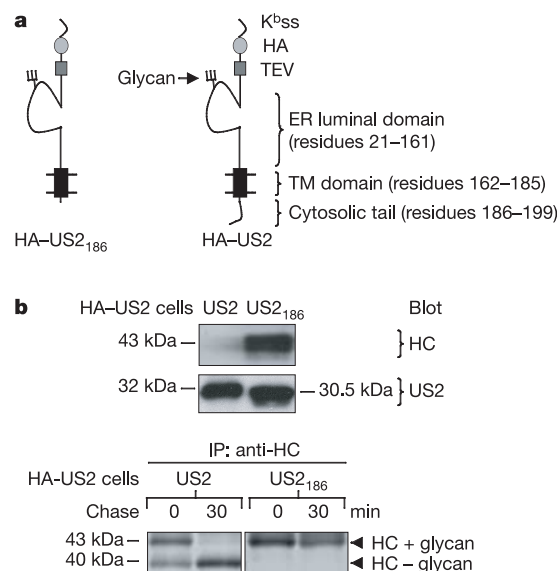


Figure 1 | Characterization of constructs used for a functional screen for US2-associated proteins. **a**, Schematic representation of US2 constructs used for affinity purification. K^bss, murine H2-K^b signal sequence; HA–US2, HA–TEV-tagged-wild type US2; US2₁₈₆, HA–TEV-tagged-tailless US2. **b**, Tailless US2 (HA–US2₁₈₆) is dislocation-incompetent. Upper panel: HC stability in U373-MG astrocytoma cells expressing tagged US2 constructs. Immunoblot analysis was performed with the indicated antibodies. Bottom panel: analysis of HC dislocation by a 15-min pulse followed by a chase in the presence of the proteasome inhibitor ZL3VS. The positions of the glycosylated and deglycosylated HC are indicated. IP, immunoprecipitation.

¹Department of Pathology and ²Pathology Functional Proteomics Center, Harvard Medical School, Boston, Massachusetts 02115, USA. ³Department of Microbiology, Mount Sinai School of Medicine, New York, New York 10021, USA. [†]Present addresses: Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA (J.L., E.S., H.L.P.); Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA (B.N.L.).

the retrieved material was then treated with tobacco etch virus (TEV) protease to release US2 and associated proteins. Mass spectrometry was used to identify polypeptides retrieved uniquely by HA-US2 and revealed, among other proteins, the prominent presence of SPP (Fig. 2a, and Supplementary Fig. 2). SPP is an ER-resident presenilin-type aspartic protease^{13,14}. The presenilin (PS)/SPP-like (SPPL) intramembrane-cleaving aspartic proteases, of which seven related members are present in the human genome¹⁵, cleave substrate polypeptides within a transmembrane region and possess two active-site aspartate residues within the conserved motifs YD and LGLGD (underlined in Fig. 2a) in adjacent membrane-spanning regions^{14,15}. In humans, SPP generates signal-peptide-derived ligands for the non-classical HLA-E molecule¹⁶ and is involved in processing of the core protein of hepatitis C virus¹⁷. It is also proposed to modulate the interaction of signal peptide remnants of HIV gp160 and prolactin with calmodulin¹⁸ and has been found in association with a truncated version of a polytopic ER protein in an *in vitro* system¹⁹.

SPP is a 40–45-kDa protein predicted to span the ER membrane seven to nine times¹³ and migrates as a 85–95-kDa SDS-resistant homodimer in SDS–polyacrylamide-gel electrophoresis (SDS–PAGE) and immunoblots; on heating or treatment with acid, the dimer dissociates and the SPP monomers can be resolved²⁰. Immunoprecipitation from HA-US2 or HA-US2₁₈₆ cells with anti-HA antibody recovered US2 (US2 blot, Fig. 2b), and immunoblotting with

anti-SPP antiserum confirmed the specific interaction of SPP with wild-type US2 but not with US2₁₈₆ (SPP blot, Fig. 2b). Conversely, US2, but not US2₁₈₆, co-immunoprecipitates with Myc-tagged SPP (Fig. 2c). We conclude that US2 interacts with SPP, an interaction that requires an intact US2 cytoplasmic tail.

Although identification of SPP in association only with functional US2 indicates a possible function for SPP in HC dislocation, we sought to obtain direct functional evidence for SPP involvement through the use of short hairpin RNAs (shRNAs) to reduce the expression of SPP²¹. A robust knockdown of SPP was achieved with three different SPP shRNA constructs, namely SPP.1, SPP.2 and SPP.3, when compared with a control (ctrl) shRNA directed against the unrelated EGFP protein⁸ (SPP blot, Fig. 2d). Whereas knockdown of SPP had no effect on HC levels in control U373 cells (data not shown), US2 cells that express each of the three SPP shRNAs have increased levels of HC (Fig. 2d, HC blot). The use of a mismatched shRNA control (mismatched SPP.1 or MM.1), as well as titration of the shRNA, confirm specificity of the knockdown attained (Supplementary Fig. 3). We conclude that the interaction of SPP with US2 is essential for US2-mediated ER dislocation of class I MHC HC molecules.

Whereas US2/SPP knockdown cells have elevated HC levels, the same decrease in SPP expression has no effect on the HC levels in US11-expressing cells (Fig. 3a). The analysis performed for US11 that led to the identification of Derlin-1 did not identify SPP⁴, and similarly, we do not retrieve SPP in US11 immunoprecipitates, or Derlin-1 in US2 immunoprecipitates (Fig. 3b). These data further support the notion that the ER degradation pathways co-opted by the HCMV immunoevasins US2 and US11 are distinct⁴.

We then addressed whether the transmembrane domain of US2 (US2 TMD) is also required for SPP interaction. We installed the K^{ss}-HA-TEV tag on a US2/CD4/US2 chimaeric construct consisting of the US2 luminal domain, the transmembrane segment of the unrelated type I membrane protein CD4, and the US2 cytosolic

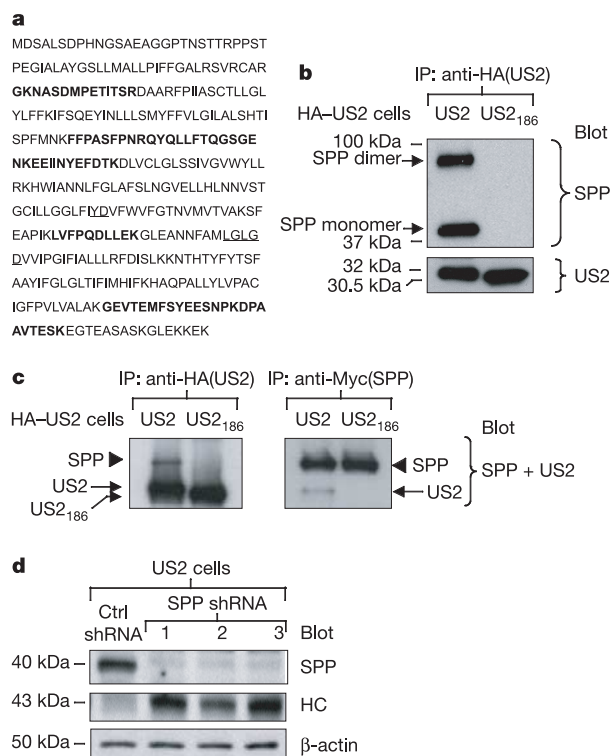


Figure 2 | SPP associates with the US2 cytosolic tail and is required for HC dislocation. **a**, SPP amino-acid sequence; peptides identified by MS/MS are shown in bold. **b**, Only functional US2 (+tail) recruits SPP. Immunoprecipitation (IP) with anti-HA antibody from native lysates of HA-US2 or HA-US2₁₈₆ cells. Subsequent immunoblot analysis with the indicated antibodies. **c**, SPP distinguishes between US2 and tailless US2₁₈₆. Immunoprecipitation with anti-HA or anti-Myc antibody from native lysates of HA-US2 or HA-US2₁₈₆ cells expressing N-Myc-tagged SPP and subsequent immunoblot analysis. A similar association was seen for monomeric and dimeric SPP (Supplementary Fig. 6b). **d**, Reduction of SPP levels by RNA-mediated interference impairs HC dislocation in US2 cells. Immunoblot analysis of SPP levels and HC stability in US2 cells expressing control (ctrl) or SPP shRNAs 1, 2 or 3. β -Actin levels serve as a loading control. Henceforth samples were boiled extensively before being loaded, so that only the SPP monomer was visible (Supplementary Fig. 6a).

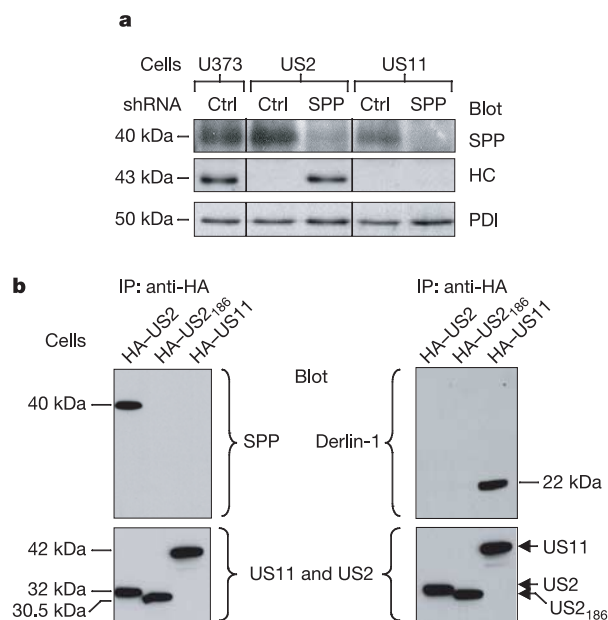


Figure 3 | The SPP-dependent ER dislocation pathway co-opted by US2 is distinct from that used by US11. **a**, SPP knockdown impairs HC dislocation by US2 but not US11. Immunoblot analysis of SPP levels and HC stability in US2 or US11 cells expressing control (ctrl) or SPP (1 + 2 + 3) shRNAs. Protein disulphide isomerase (PDI) levels were used as a loading control. U373 cells serve as a control for HC levels in the absence of US2 and US11. **b**, US2 binds SPP and not Derlin-1; US11 binds Derlin-1 and not SPP. Results are from immunoprecipitation (IP) with anti-HA antibody from native lysates of HA-US2, HA-US2₁₈₆ and HA-US11 cells followed by immunoblot analysis.

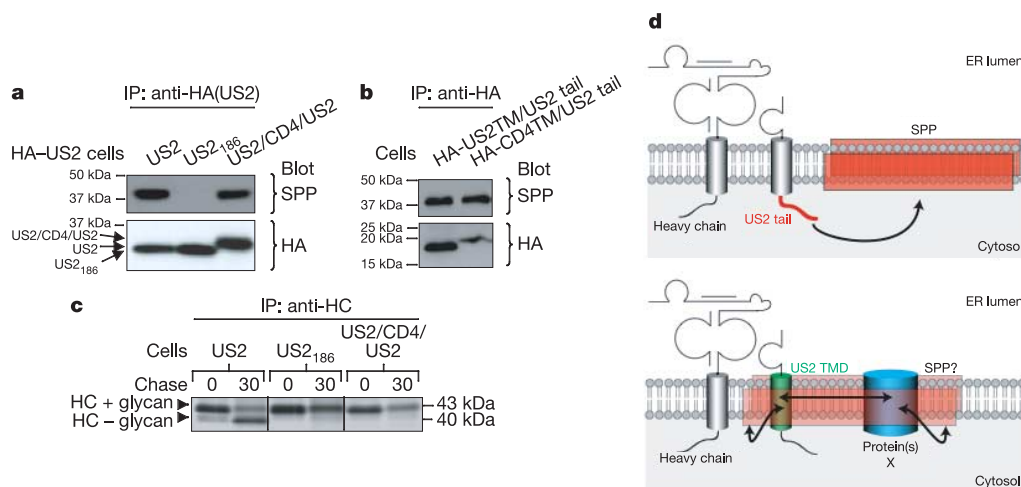


Figure 4 | US2-mediated dislocation requires US2 tail-dependent SPP recruitment and additional US2 TMD-dependent interactions. **a**, The US2 TMD is dispensable for SPP association. Results are from immunoprecipitation (IP) with anti-HA antibody followed by immunoblot analysis with the antibodies indicated. **b**, The US2 tail is not only necessary but also sufficient for SPP association. Results are from immunoprecipitation (IP) with anti-HA antibody followed by immunoblot

analysis with the antibodies indicated. **c**, The US2 transmembrane domain is critical for dislocation. Analysis of HC dislocation by a 15-min pulse followed by a chase in the presence of the proteasome inhibitor ZLVS. The positions of the glycosylated and deglycosylated HC are indicated. **d**, **e**, Hypothetical model for the contribution of the US2 tail (**d**) and the TMD (**e**) to the dislocation of HC from the ER membrane.

tail (Supplementary Figs 4 and 5a). Because US2/CD4/US2 is able to bind SPP (Fig. 4a), the US2 TMD must be dispensable for interaction with SPP. In fact, a construct composed of the K^bss-HA-TEV tag, the CD4 transmembrane domain and the US2 tail (HA-CD4TM/US2 tail) still binds SPP (Fig. 4b), so the US2 tail is not only necessary but also sufficient for association with SPP.

Interestingly, the US2/CD4/US2 chimera, although maintaining the ability to bind HC (Supplementary Fig. 5b) and SPP, fails to catalyse HC dislocation, because no deglycosylated HC intermediate is detected (Fig. 4c). HC dislocation is therefore dependent not just on (US2 tail-mediated) recruitment of SPP but also on additional (US2 TMD-mediated) interactions within the plane of the membrane (Figs 4d, e). The US2 TMD might be involved in the further engagement of SPP or alternatively in the recruitment or engagement of other protein(s) involved in dislocation. Thus, although SPP is unlikely to be the sole host-derived component of the US2 pathway, it is nonetheless essential. We have yet to explore the proteins other than SPP that are differentially associated with functional and non-functional US2 (Supplementary Fig. 2). At least some of these are also likely to have a function in dislocation.

Our work provides evidence for a function of the intramembrane-cleaving protease SPP in ER dislocation of class I MHC molecules. This raises the question of the mechanism of its involvement. We do not know whether the structural and/or catalytic properties of SPP itself or a post-targeting function of a SPP substrate are required for its function in dislocation²². Interaction of US2 with SPP possibly occurs through the recognition of a stretch of hydrophobic amino-acid residues on the US2 tail. Because US2 is a type I membrane glycoprotein, the topology of its transmembrane and tail segments is opposite to that of signal peptides released by signal peptidase, which are the substrates for SPP. Although binding of US2 to SPP might take place by virtue of the generally hydrophobic character of the US2 tail, intramembrane cleavage of the US2 TMD by SPP in this inverted orientation would presumably not occur, as seen for other proteases^{23,24}. Indeed, substrate recruitment and subsequent cleavage by SPP may be separable events, as shown for the related presenilins^{25,26}. SPP-mediated intramembrane cleavage within the HC TMD is inconsistent with the observed recovery of full-length HC in the dislocation reaction^{2,3,27,28}. Alternatively, US2 binding to SPP could modulate its activity and/or structural properties. A putative link between SPP and the unfolded protein response also remains to be explored²⁹.

Although the removal of signal peptide remnants from the ER membrane is a function assigned to SPP in animals and plants, a gene that encodes an orthologue of this enzyme is absent from the yeast genome^{14,22}. The function of SPP in higher eukaryotes might therefore not be limited to signal peptide processing and might extend to processes such as protein dislocation from the ER. Other member(s) of the PS/SPPL superfamily of intramembrane-cleaving aspartic proteases, so far of unknown function¹⁵, might similarly be involved in the disposal of different degradation substrates. Because of its interaction with a truncated opsin chain *in vitro*, SPP was implicated in the recognition of misassembled transmembrane domains during membrane-protein quality control at the ER¹⁹. Our data link SPP to dislocation and indicate that US2 might even be uncovering a more general connection between regulated intramembrane proteolysis³⁰ and quality control in the ER.

METHODS

Cell culture. U373-MG astrocytoma/glioblastoma cells were cultured as described¹¹. U373-MG cells stably expressing HA-US2, HA-US2₁₈₆, HA-US2/CD4/US2, HA-US2TM/US2 tail and HA-CD4 TM/US2 tail were obtained by retroviral transduction⁴ with vectors constructed with the pLNCX (Clontech) backbone and selected and maintained in DMEM medium containing 0.5 mg ml⁻¹ geneticin (Invitrogen). US2, US11 and HA-US11 cells have been described previously^{4,27}. Myc-tagged SPP-expressing cells were obtained by retroviral transduction with SPP in pLHCX retroviral vector (Clontech) and selected and maintained in DMEM containing 0.250 mg ml⁻¹ hygromycin B (Roche).

cDNA constructs. An N-terminal K^bss-HA-TEV tag followed by a small spacer⁴ was fused to US2 residues 21–199 (US2) or 21–186 (US2₁₈₆)¹² for the large-scale affinity purification. The same tag was later added to US2/CD4/US2, US2TM/US2 tail and CD4TM/US2 tail. Human SPP cDNA¹⁴ was obtained from B. Martoglio, amplified by polymerase chain reaction, N- and C-terminally Myc-tagged and subcloned into the pLHCX retroviral vector (Clontech). The control (ctrl) enhanced green fluorescent protein (EGFP) shRNA construct has been described previously⁸. SPP shRNA target sequences for pRETRO-SUPER expression²¹ were chosen with an algorithm available at <http://jura.wi.mit.edu/siRNAext/> and were as follows: 5'-AAGGCCTCGAAGCAAACAAC-3' (SPP.1), 5'-AGGGAGAAGTGACAGAGATG-3' (SPP.2) and 5'-AAGAATGCTTCAGACATGCTC-3' (SPP.3). shRNA-expressing cells were obtained by retroviral transduction^{8,21} and assayed 48–96 h after infection with shRNA-expressing retroviruses. In addition to the unrelated EGFP shRNA control, we made an additional RNA-mediated interference mismatched control by making a two-base-pair change in the middle of the SPP.1 shRNA (mismatched SPP.1 or MM.1: 5'-AAGGCCTCGAGGCGAACAAC-3') and showed that the SPP.1

shRNA effect was ablated in cells expressing MM.1 (Supplementary Fig. 3).

Pulse-chases. In brief, cells were starved for methionine and cysteine for 1 h in the presence of the proteasome inhibitor ZL3VS, pulsed with radioactive ^{35}S -labelled methionine/cysteine (Perkin Elmer) for 15 min, and then chased with an excess of non-radioactive amino acids for 0 and 30 min. Lysates in 0.5% Nonidet P40 were then equalized for incorporation of radioactive material with counts in the trichloroacetic acid precipitate and immunoprecipitated with the indicated antibodies. Immunoprecipitates were subject to 10% or 12.5% SDS-PAGE and fluorography as described previously^{2,3} (Figs 1c and 4c, and Supplementary Figs 1a, b and 5b). Digestions with endoglycosidase H were performed, where indicated, in accordance with the manufacturer's instructions (New England BioLabs).

Immunoprecipitations and western blots. In brief, lysates in 1% digitonin were normalized for total protein content with the bicinchoninic acid assay (Pierce). Typically, 150–250 μg of total cellular protein was used for immunoprecipitation with anti-HA or anti-Myc antibody, subjected to 4–15% (Fig. 2b, c, and Supplementary Fig. 6b) or 10% (Figs 3b and 4a, b) SDS-PAGE, transferred to a poly(vinylidene difluoride) (PVDF) membrane and immunoblotted with the antibodies indicated.

Western blots. Cells were lysed in 0.5% Nonidet P40 and total protein content was normalized with the bicinchoninic acid assay (Pierce). Typically, 10–20 μg of total cellular protein was loaded on each lane of a 4–15% SDS-PAGE gel (Figs 1b, 2d and 3a, and Supplementary Fig. 3), transferred to a PVDF membrane and immunoblotted with the antibodies indicated.

Antibodies. Rabbit anti-human HC, anti-US2 and anti-human PDI antibodies and mouse W6/32 antibody have been described previously¹¹. Anti-Myc antibodies were from Invitrogen, anti-SPP antibodies were from AbCam, anti-HA (3F10 clone) antibodies were from Roche and anti- β -actin antibodies were from Sigma.

Large-scale affinity purification and mass spectrometry. The procedure was adapted from one described previously⁴. In brief, astrocytoma cells (10^8) stably expressing HA-US2, HA-US2₁₈₆, or no tagged protein (control cells), were lysed in 30 ml of ice-cold lysis buffer (1% digitonin in 100 mM Tris-HCl pH 7.4, 20 mM NaCl, 70 mM KCl, 12 mM MgCl₂ containing 1 mM PMSF and complete protease inhibitors) with rocking at 4 °C for 1 h. The lysate was cleared of debris by centrifugation at 20,000 g for 15 min. US2 and associated proteins were retrieved by immunoprecipitation with 250 μl of compacted anti-HA antibody beads from 10 mg of cleared lysate. After incubation for 2 h, beads were extensively washed in wash buffer (composition as lysis buffer, except with 0.2% digitonin and without protease inhibitors). Bound material was eluted by treatment with 50 U TEV protease (Invitrogen) at 30 °C for 2 h in TEV buffer with 1 mM dithiothreitol and 0.2% digitonin. The eluate was boiled in reducing conditions for 3 min, and the resulting polypeptides were separated by 8% SDS-PAGE and revealed by silver staining. The bands of interest were excised, subjected to trypsinolysis, separated by liquid chromatography, analysed by MS/MS and database-searched as described⁴. The data obtained are summarized in Supplementary Fig. 2.

Received 9 January; accepted 18 April 2006.

Published online 31 May 2006.

- Tortorella, D., Gewurz, B. E., Furman, M. H., Schust, D. J. & Ploegh, H. L. Viral subversion of the immune system. *Annu. Rev. Immunol.* **18**, 861–926 (2000).
- Wiertz, E. J. et al. Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. *Nature* **384**, 432–438 (1996).
- Wiertz, E. J. et al. The human cytomegalovirus US11 gene product dislocates MHC class I heavy chains from the endoplasmic reticulum to the cytosol. *Cell* **84**, 769–779 (1996).
- Lilley, B. N. & Ploegh, H. L. A membrane protein required for dislocation of misfolded proteins from the ER. *Nature* **429**, 834–840 (2004).
- Elgaard, L. & Helenius, A. Quality control in the endoplasmic reticulum. *Nature Rev. Mol. Cell Biol.* **4**, 181–191 (2003).
- Meusser, B., Hirsch, C., Jarosch, E. & Sommer, T. ERAD: the long road to destruction. *Nature Cell Biol.* **7**, 766–772 (2005).
- Gewurz, B. E., Wang, E. W., Tortorella, D., Schust, D. J. & Ploegh, H. L. Human cytomegalovirus US2 endoplasmic reticulum-lumenal domain dictates association with major histocompatibility complex class I in a locus-specific manner. *J. Virol.* **75**, 5197–5204 (2001).
- Lilley, B. N. & Ploegh, H. L. Multiprotein complexes that link dislocation, ubiquitination, and extraction of misfolded proteins from the endoplasmic reticulum membrane. *Proc. Natl Acad. Sci. USA* **102**, 14296–14301 (2005).
- Ye, Y., Meyer, H. H. & Rapoport, T. A. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature* **414**, 652–656 (2001).
- Ye, Y., Shibata, Y., Yun, C., Ron, D. & Rapoport, T. A. A membrane protein complex mediates retro-translocation from the ER lumen into the cytosol. *Nature* **429**, 841–847 (2004).
- Gewurz, B. E., Ploegh, H. L. & Tortorella, D. US2, a human cytomegalovirus-encoded type I membrane protein, contains a non-cleavable amino-terminal signal peptide. *J. Biol. Chem.* **277**, 11306–11313 (2002).
- Furman, M. H., Ploegh, H. L. & Tortorella, D. Membrane-specific, host-derived factors are required for US2- and US11-mediated degradation of major histocompatibility complex class I molecules. *J. Biol. Chem.* **277**, 3258–3267 (2002).
- Friedmann, E. et al. Consensus analysis of signal peptide peptidase and homologous human aspartic proteases reveals opposite topology of catalytic domains compared with presenilins. *J. Biol. Chem.* **279**, 50790–50798 (2004).
- Weihofen, A., Binns, K., Lemberg, M. K., Ashman, K. & Martoglio, B. Identification of signal peptide peptidase, a presenilin-type aspartic protease. *Science* **296**, 2215–2218 (2002).
- Martoglio, B. & Golde, T. E. Intramembrane-cleaving aspartic proteases and disease: presenilins, signal peptide peptidase and their homologs. *Hum. Mol. Genet.* **12** (spec. no. 2), R201–R206 (2003).
- Lemberg, M. K., Bland, F. A., Weihofen, A., Braud, V. M. & Martoglio, B. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. *J. Immunol.* **167**, 6441–6446 (2001).
- McLauchlan, J., Lemberg, M. K., Hope, G. & Martoglio, B. Intramembrane proteolysis promotes trafficking of hepatitis C virus core protein to lipid droplets. *EMBO J.* **21**, 3980–3988 (2002).
- Martoglio, B., Graf, R. & Dobberstein, B. Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *EMBO J.* **16**, 6636–6645 (1997).
- Crawshaw, S. G., Martoglio, B., Meacock, S. L. & High, S. A misassembled transmembrane domain of a polytopic protein associates with signal peptide peptidase. *Biochem. J.* **384**, 9–17 (2004).
- Nyborg, A. C. et al. Signal peptide peptidase forms a homodimer that is labeled by an active site-directed gamma-secretase inhibitor. *J. Biol. Chem.* **279**, 15153–15160 (2004).
- Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
- Martoglio, B. Intramembrane proteolysis and post-targeting functions of signal peptides. *Biochem. Soc. Trans.* **31**, 1243–1247 (2003).
- Roques, B. P., Lucas-Soroca, E., Chaillet, P., Costentin, J. & Fournie-Zaluski, M. C. Complete differentiation between enkephalinase and angiotensin-converting enzyme inhibition by retro-thiorphan. *Proc. Natl Acad. Sci. USA* **80**, 3178–3182 (1983).
- Tarasova, N. I. et al. Transmembrane inhibitors of P-glycoprotein, an ABC transporter. *J. Med. Chem.* **48**, 3768–3775 (2005).
- Kornilova, A. Y., Das, C. & Wolfe, M. S. Differential effects of inhibitors on the gamma-secretase complex. Mechanistic implications. *J. Biol. Chem.* **278**, 16470–16473 (2003).
- Lemberg, M. K. & Martoglio, B. On the mechanism of SPP-catalysed intramembrane proteolysis; conformational control of peptide bond hydrolysis in the plane of the membrane. *FEBS Lett.* **564**, 213–218 (2004).
- Blom, D., Hirsch, C., Stern, P., Tortorella, D. & Ploegh, H. L. A glycosylated type I membrane protein becomes cytosolic when peptide: N-glycanase is compromised. *EMBO J.* **23**, 650–658 (2004).
- Misaghi, S., Pacold, M. E., Blom, D., Ploegh, H. L. & Korbel, G. A. Using a small molecule inhibitor of peptide: N-glycanase to probe its role in glycoprotein turnover. *Chem. Biol.* **11**, 1677–1687 (2004).
- Shaffer, A. L. et al. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity* **21**, 81–93 (2004).
- Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100**, 391–398 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. High for reagents, A. Weihofen and M. Wolfe for technical advice and insight, A. Weihofen and M. Lemberg for critical reading of the manuscript, T. DiCesare for Fig. 4c, and members of the Ploegh laboratory for discussions. J.L. is a Gulbenkian PhD Program Fellow with a grant from the Portuguese Science and Technology Foundation (FCT). B.N.L. was supported by a Howard Hughes Medical Institute Predoctoral Fellowship. This work was supported by grants from the NIH to H.L.P.

Author Contributions B.N.L. designed the HA-US2 and HA-US2₁₈₆ constructs, supplied the retroviral supernatants for transduction of astrocytoma cells and provided critical reading of the manuscript. V.N. and D.T. designed and supplied the untagged US2-CD4-US2 cDNA. All other constructs were designed by J.L. Mass spectrometry analysis was performed by E.S. All experiments were designed and performed by J.L. Data analysis, interpretation and writing of the manuscript were by J.L. and H.L.P.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.L.P. (ploegh@wi.mit.edu).

LETTERS

CRD-BP mediates stabilization of β TrCP1 and c-myc mRNA in response to β -catenin signalling

Felicite K. Noubissi¹, Irina Elcheva¹, Neehar Bhatia¹, Abbas Shakoori², Andrei Ougolkov^{2†}, Jianghuai Liu³, Toshinari Minamoto², Jeff Ross⁴, Serge Y. Fuchs³ & Vladimir S. Spiegelman¹

Although constitutive activation of β -catenin/Tcf signalling is implicated in the development of human cancers¹, the mechanisms by which the β -catenin/Tcf pathway promotes tumorigenesis are incompletely understood. Messenger RNA turnover has a major function in regulating gene expression and is responsive to developmental and environmental signals. mRNA decay rates are dictated by *cis*-acting elements within the mRNA and by *trans*-acting factors, such as RNA-binding proteins (reviewed in refs 2, 3). Here we show that β -catenin stabilizes the mRNA encoding the F-box protein β TrCP1, and identify the RNA-binding protein CRD-BP (coding region determinant-binding protein) as a previously unknown target of β -catenin/Tcf transcription factor. CRD-BP binds to the coding region of β TrCP1 mRNA. Overexpression of CRD-BP stabilizes β TrCP1 mRNA and elevates β TrCP1 levels (both in cells and *in vivo*), resulting in the activation of the Skp1-Cullin1-F-box protein (SCF) ^{β TrCP} E3 ubiquitin ligase and in accelerated turnover of its substrates including I κ B and β -catenin. CRD-BP is essential for the induction of both β TrCP1 and c-Myc by β -catenin signalling in colorectal cancer cells. High levels of CRD-BP that are found in primary human colorectal tumours exhibiting active β -catenin/Tcf signalling implicates CRD-BP induction in the upregulation of β TrCP1, in the activation of dimeric transcription factor NF- κ B and in the suppression of apoptosis in these cancers.

Physiological and pathophysiological consequences of activated β -catenin/Tcf signalling are mediated by multiple target genes, most of which are direct Tcf-dependent transcriptional targets (reviewed in refs 1, 4). In contrast, induction of β TrCP1 expression by β -catenin is independent of β TrCP1 gene transcription⁵. We found that activation of the β -catenin signalling pathway by the expression of a stabilized β -catenin^{S33Y} mutant drastically increased the half-life of β TrCP1 mRNA (Fig. 1a). This β TrCP1 mRNA stabilization was abolished by co-expression of dnTcf4 (Fig. 1b), indicating that observed phenomenon might depend on Tcf-driven transcription. These results show that activation of β -catenin/Tcf signalling leads to stabilization of β TrCP1 mRNA.

Stabilization of β TrCP1 mRNA could arise from Tcf-mediated induction of an mRNA-stabilizing protein. Most such proteins interact with known determinants of mRNA stability that are located within the 3'-untranslated region (UTR)^{2,3}. Although removal of the 3'-UTR stabilized β TrCP1 mRNA, fusion of this 3'-UTR to an unrelated mRNA did not confer responsiveness to β -catenin (data not shown). β -Catenin/Tcf induced the binding of roughly 70-kDa protein to β TrCP1 mRNA (Supplementary Fig. S1a and Supplementary Information). The protein binding site(s) were located within the mRNA coding region, because the fragment of β TrCP1 mRNA

coding region from base pairs (bp) 577–1182 was sufficient for the interaction with this protein (Supplementary Fig. S1a, lanes 3 and 4). This segment of β TrCP1 mRNA inserted in-frame into the β -globin coding sequence ('GBG') conferred instability on a stable mRNA ('GGG') and the destabilization effect was attenuated after activation of the β -catenin signalling pathway (Fig. 1c). We conclude that the coding region 577–1182 bp of β TrCP1 mRNA contains a *cis*-acting factor that is both necessary and sufficient to confer the regulation: it can confer instability on a stable mRNA and the destabilization effect is abolished after activation of β -catenin/Tcf signalling.

A β -catenin-induced 70-kDa protein that could be purified by virtue of its binding to the β TrCP1 mRNA fragment 577–1182 bp was identified as CRD-BP (Fig. 2a, and Supplementary Fig. S1b). Furthermore, Flag-tagged CRD-BP expressed in cells was found to interact with the 577–1182 bp RNA substrate (Fig. 2b). The specificity of the binding reaction was confirmed by an ultraviolet crosslinking/immunoprecipitation experiment (Fig. 2c) as well as by binding assays with CRD-BP expressed in cells (Supplementary Fig. S1c) or translated *in vitro* (Supplementary Fig. S1d). We conclude that CRD-BP (which is upregulated in response to β -catenin activation) interacts with the 577–1182 bp fragment of the coding region of β TrCP1 mRNA. CRD-BP, which is also known as insulin-like growth factor 2 (IGF2) mRNA-binding protein (IMP-1), is a multifunctional RNA-binding protein that recognizes *c-myc*, *IGF2*, *tau* and β -actin mRNAs and H19 RNA^{6–12}.

Overexpression of CRD-BP in cells led to stabilization of β TrCP1 mRNA (Fig. 2d) and its accumulation (Supplementary Fig. S1e). CRD-BP induced β TrCP1 expression *in vivo*. Transgenic WAP-CRD-BP female mice express high levels of CRD-BP in mammary tissue during pregnancy. β TrCP1 overexpression is also observed in the same samples, establishing a strong link between overexpression of CRD-BP and β TrCP1 (Fig. 2e). We have previously shown that WAP-CRD-BP mice develop spontaneous mammary tumours¹³. Conversely, transgenic mice that express high levels of β TrCP1 in mammary epithelium (murine-mammary-tumour virus (MMTV)- β TrCP) display a hyperplastic phenotype and also develop tumours¹⁴. It is therefore plausible that overexpression of β TrCP1 in WAP-CRD-BP mice contributes to their tumorigenic phenotype.

Numerous reports indicate that levels of β TrCP determine the stability of its substrates under the conditions of constitutive phosphorylation¹⁵. Consistent with these reports is our observation that ectopic expression of CRD-BP not only led to elevated levels of β TrCP1 but also accelerated the degradation of known β TrCP1 substrates such as β -catenin and I κ B α (Fig. 2f). Because neither I κ B kinase activity nor I κ B/ β -catenin phosphorylation was increased by CRD-BP (Supplementary Fig. S2a, and data not shown), it is likely

¹Department of Dermatology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin 53706, USA. ²Division of Translational and Clinical Oncology, Cancer Research Institute, Kanazawa University, Kanazawa 920-0934, Japan. ³Department of Animal Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁴McArdle Laboratory for Cancer Research, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin 53706, USA. [†]Present address: Division of Oncology Research, Mayo Clinic, Rochester, Minnesota 55905, USA.

that rapid proteolysis of these substrates might be attributed to CRD-BP-induced levels of β TrCP1 and activation of $SCF^{\beta\text{TrCP1}}$ E3 ligase activity. Indeed, this effect of CRD-BP was abrogated by β TrCP1 knockdown (Fig. 2f). CRD-BP therefore seems to be the first regulator of mRNA stability that is also important in regulating the stability of specific protein substrates.

β -Catenin induces the expression of CRD-BP (Fig. 3a, b) but not of its homologues, IMP-2 and IMP-3 (Fig. 3a). This upregulation was dependent on Tcf (Fig. 3b) and was mediated by increased CRD-BP transcription (as measured by nuclear run-on; Fig. 3c). Transcriptional upregulation of CRD-BP by β -catenin/Tcf signalling was comparable to that of *Tcf1*, a well-known transcriptional target of β -catenin/Tcf¹⁶. We also found that the promoter region of CRD-BP efficiently interacts with β -catenin in a chromatin immunoprecipitation assay (Fig. 3d). The binding of this region (which contains a putative Tcf-binding site) to β -catenin is induced by activation of the β -catenin signalling pathway by either the expression of

β -catenin^{S33Y}/Tcf4 (Fig. 3e) or by treatment of cells with recombinant Wnt3A protein (Fig. 3f), indicating that CRD-BP might be a primary target gene for β -catenin/Tcf bipartite transcription factor.

Knockdown of endogenous CRD-BP by specific short hairpin RNA (shRNA; characterized in Supplementary Fig. S2b, c) blocks the β -catenin/Tcf-dependent induction of β TrCP1 expression (Fig. 4a), prevents the stabilization of endogenous β TrCP1 mRNA (Supplementary Fig. S2d) and decreases the levels of β TrCP1 mRNA in

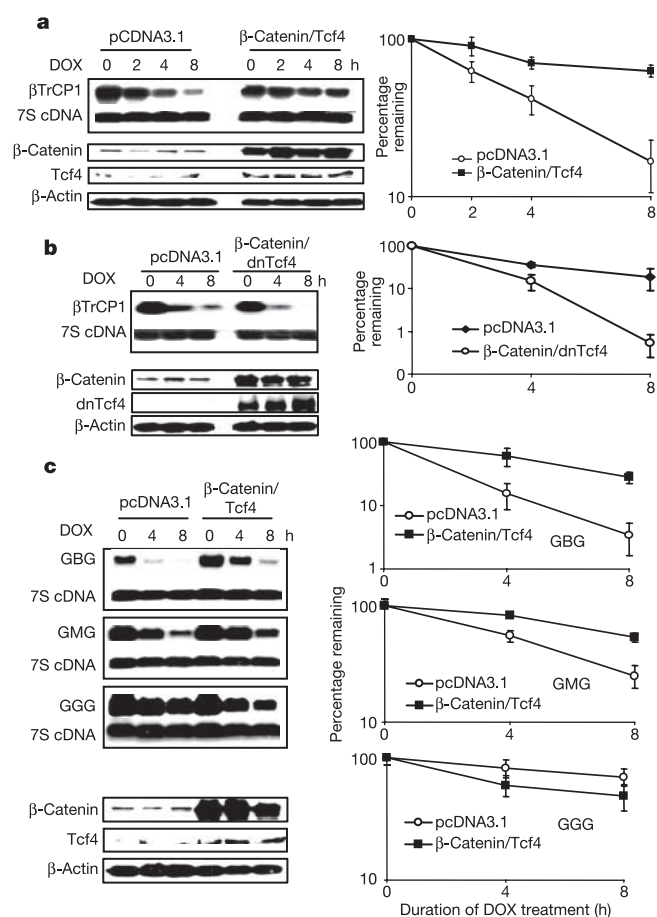


Figure 1 | β -catenin/Tcf signalling stabilizes β TrCP1 mRNA. **a**, 293T cells were co-transfected with Tet-off, pBI-G- β TrCP1 and pcDNA3.1 or β -catenin/Tcf4. Transcription was stopped by treatment with doxycycline (DOX) for the indicated durations, and the stability of the indicated mRNAs was analysed by northern blotting (left panels) and presented graphically (right panels). Values presented are % ratios of the amounts of β TrCP1 mRNA remaining at different time points to the initial (time point '0') quantity. Data are means $\pm$ s.d. of three or four separate experiments. Immunoblot analysis of indicated proteins is presented in the lower panels. **b**, 293T cells were co-transfected with Tet-off, pBI-G- β TrCP1 and pcDNA3.1 or β -catenin/dominant negative Tcf4 (dnTcf4) and analysed as in **a**. **c**, Fusions of β -globin with glyceraldehyde-3-phosphate dehydrogenase ('GGG'), with the 577–1182 bp fragment of β TrCP1 ('GBG') or with the coding region determinant of c-Myc ('GMG') were expressed in 293T cells under the control of a Tet-off system along with the indicated constructs, and analysed as in **a**.

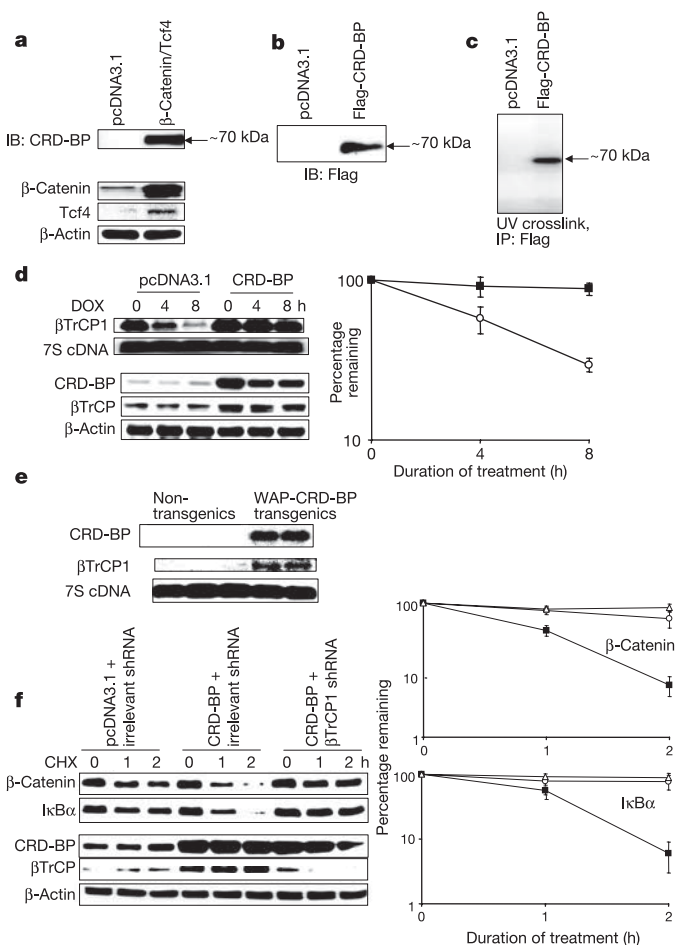


Figure 2 | CRD-BP binds to β TrCP1 mRNA, stabilizes it and induces β TrCP1 expression and activities. **a**, Immunoblot analysis of proteins processed as described in the legend to Supplementary Fig. 1b with polyclonal anti-CRD-BP antibody. The lower panel shows an immunoblot analysis of indicated proteins isolated from a whole-cell protein extracts. **b**, Immunoblot (IB) analysis of RSW proteins affinity-purified on a β TrCP1 mRNA (577–1182 bp) column isolated from 293T cells transfected as indicated. **c**, Flag immunoprecipitation (IP) of ultraviolet-crosslinked complexes of the 577–1182 bp fragment of β TrCP1 mRNA and RSW from 293T cells transfected as indicated. **d**, The stability of β TrCP1 mRNA in 293T cells expressing CRD-BP as described in the legend to Fig. 1a was analysed by northern blotting (upper panels) and presented graphically (right panel). Filled squares, CRD-BP; open circles, pcDNA3.1. An immunoblot analysis of the indicated proteins is presented in the lower panel on the left. **e**, Northern blot analysis of RNA isolated from mammary tissues of 12-day pregnant non-transgenic and transgenic WAP-CRD-BP mice as indicated. **f**, The stability of β -catenin and I κ B α was analysed by immunoblot of protein extracts isolated from 293T cells transfected as indicated and treated with cycloheximide (CHX) for the indicated durations. Open circles, pcDNA3.1 plus irrelevant shRNA; filled squares, CRD-BP plus irrelevant shRNA; open triangles, CRD-BP plus β TrCP1 shRNA. Values presented are % ratios of the amounts of β -catenin or I κ B α remaining at different time points to the initial (time point '0') quantity. Data are means $\pm$ s.d. of three separate experiments.

human colorectal cancer cells that exhibit constitutive activation of β -catenin/Tcf signalling pathway (Fig. 4b, and Supplementary Fig. S3b). These results confirm the role of CRD-BP in the regulation of β TrCP1 expression by β -catenin. We propose a model for a negative-feedback regulation: β -catenin/TCF activates CRD-BP and CRD-BP in turn stabilizes mRNA of β TrCP1, whose protein product degrades β -catenin pending its efficient phosphorylation within the destruction motif. Indeed, our results show the transient nature of Wnt3A-induced β -catenin–DNA interaction that can be significantly prolonged by CRD-BP knockdown (Supplementary Fig. S3a).

CRD-BP function seems to be important for the pathogenesis of human colorectal cancer, a disease in which the activation of β -catenin/Tcf signalling is well established (reviewed in ref. 17). We have previously shown that upregulation of β TrCP1 by β -catenin resulted in the modulation of NF- κ B signalling by means of β TrCP-mediated degradation of I κ B in cells⁵. Furthermore, in primary human colorectal cancers, activation of β -catenin and accumulation of β TrCP1 are correlated with the nuclear localization of NF- κ B and a decreased apoptotic index¹⁸. Indeed, NF- κ B activation is known to contribute to tumour development by suppressing apoptosis in a vast majority of solid tumours (reviewed in ref. 19). We also observed a remarkable correlation between the activation of β -catenin signalling and the expression of CRD-BP, and β TrCP1 in human colorectal tumours (Fig. 4c), indicating that the induction of CRD-BP by β -catenin signalling might be an important mechanism of elevating

β TrCP1 mRNA levels in human colorectal cancers. These data are in line with our previous observation that CRD-BP is induced in most human colorectal tumours²⁰. Moreover, knockdown of CRD-BP noticeably decreased β TrCP1 expression in colorectal cancer cell lines (Fig. 4b, and Supplementary Fig. S3b). It also led to inhibition of NF- κ B activity, induction of apoptosis and suppression of colony formation in HCT116 colorectal cancer cells (Supplementary Fig. S3c–e). Another possibility is that the induction of CRD-BP by β -catenin might affect the expression of other genes whose mRNA is stabilized by CRD-BP. One such candidate is the proto-oncogene *c-myc*. CRD-BP was shown to bind to the coding region determinant (CRD) of *c-myc* mRNA and to shield the mRNA from attack by an endoribonuclease^{10–12}. Indeed, β -catenin/Tcf-induced expression of *c-Myc* was abolished by CRD-BP knockdown (Fig. 4a), despite the fact that *c-Myc* itself is a direct transcriptional target of β -catenin/Tcf signalling²¹. In addition, CRD of *c-myc* mRNA inserted in-frame into the β -globin coding sequence (GMG) enabled stabilization of the resulting chimaeric mRNA by β -catenin signalling (Fig. 1c). Colorectal cell lines with constitutively activated β -catenin signalling (HCT116 and Caco 2) express high levels of *c-Myc*. Transfection of these cells with CRD-BP shRNA noticeably reduces the expression of *c-myc* mRNA (Fig. 4b, and Supplementary Fig. S3b), in spite of a low efficiency of transfection in these cells. These data indicate that stabilization of *c-myc* mRNA by CRD-BP is probably an important component of the regulation of *c-Myc* expression by β -catenin signalling.

Cancer cells of different origins often use the mechanisms of

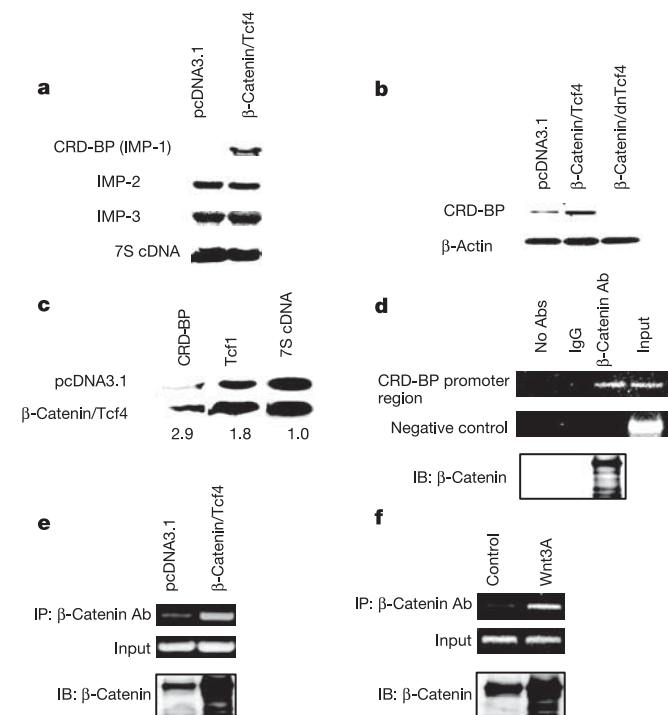


Figure 3 | β -Catenin/Tcf signalling induces expression of CRD-BP.

a, Northern blot analysis of RNA isolated from 293T cells transfected as indicated. **b**, Immunoblot analysis of protein isolated from 293T cells transfected as indicated. **c**, Nuclear run-on analysis. CRD-BP, Tcf1 or 7S cDNA probes (as indicated) hybridized with [³²P]UTP-labelled RNA from the nuclei of 293T cells, which were transfected as indicated. Numbers below the lanes reflect the fold inductions (quantified by densitometry). **d–f**, Chromatin immunoprecipitation assay of the promoter region of *CRD-BP* (–1333 to –1607). **d**, Specificity of the assay was analysed in 293T cells transfected with β -catenin/Tcf4. Ab, antibody; IB, immunoblot. Inducibility of *CRD-BP* promoter binding by β -catenin was analysed in 293T cells transfected as indicated (**e**) or in HeLa cells treated with the recombinant mouse Wnt-3A (100 ng ml⁻¹) for 2 h (**f**). IP, immunoprecipitation. Lower panels: immunoblot analysis of protein isolated from immunoprecipitates as indicated.

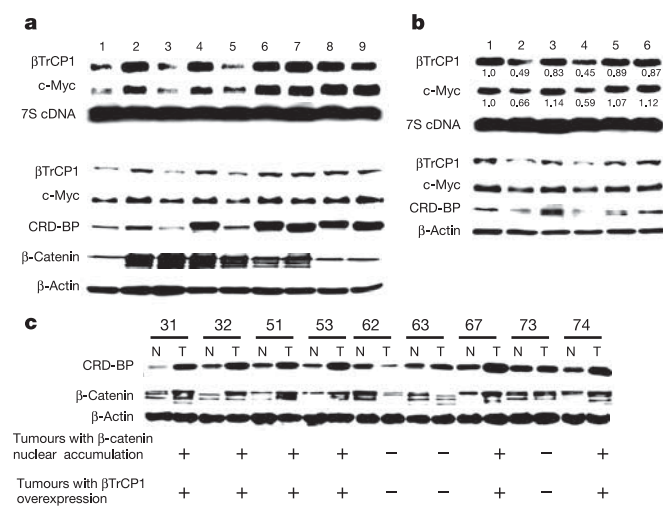


Figure 4 | CRD-BP is indispensable for the induction of β TrCP1 and *c-Myc* by β -catenin/Tcf signalling.

a, Northern blot (upper panel) and immunoblot (lower panel) analyses on 293T cells transfected as indicated. Lane 1, irrelevant shRNA; lane 2, β -catenin/Tcf4 plus irrelevant shRNA; lane 3, β -catenin/Tcf4 plus CRD-BP shRNA-3; lane 4, β -catenin/Tcf4 plus CRD-BP shRNA-3 plus mouse CRD-BP; lane 5, β -catenin/Tcf4 plus CRD-BP shRNA-3 plus human CRD-BP; lane 6, β -catenin/Tcf4 plus irrelevant shRNA plus mouse CRD-BP; lane 7, β -catenin/Tcf4 plus irrelevant shRNA plus human CRD-BP; lane 8, irrelevant shRNA plus mouse CRD-BP; lane 9, irrelevant shRNA plus human CRD-BP. **b**, Northern blot analysis of RNA isolated from HCT116 colorectal cells, transfected as indicated. Numbers below the lanes reflect the relative amounts of mRNA (quantified by densitometry). Lower panel: immunoblot analysis of protein isolated from cells transfected as indicated. Lane 1, irrelevant shRNA; lane 2, CRD-BP shRNA-3; lane 3, CRD-BP shRNA-3 plus mouse CRD-BP; lane 4, CRD-BP shRNA-3 plus human CRD-BP; lane 5, irrelevant shRNA plus mouse CRD-BP; lane 6, irrelevant shRNA plus human CRD-BP. **c**, Immunoblot analysis of proteins prepared from primary human colorectal tumours (T) and apparently normal colon tissues (N). Panels indicating β -catenin nuclear accumulation and β TrCP1 overexpression in tumours are described in detail elsewhere¹⁸.

normal embryonic development to achieve their malignant status, and the Wnt/ β -catenin/Tcf signalling pathway is a classic example of this. Coincidentally, CRD-BP was previously shown to be highly expressed during normal embryonic development^{20,22–25} as well as in a variety of human tumours^{20,24–27}. This oncofetal pattern of CRD-BP expression could potentially be a result of β -catenin/Tcf activation during embryogenesis and tumorigenesis. Transcriptional upregulation of CRD-BP by β -catenin/Tcf might therefore contribute to pleiotropic effects of the Wnt/ β -catenin/Tcf signalling pathway and could provide the foundation for the hypothesis that CRD-BP might have a significant function in the oncogenic effect of β -catenin. Future studies will define the role of CRD-BP in colorectal tumorigenesis induced by activated β -catenin/Tcf signalling.

METHODS

Expression vectors. We have subcloned the full-length β TrCP1 complementary DNA into two vectors: pBIG (Clontech) under the control of TRE promoter, and pcDNA3.1 (Invitrogen) downstream of the T7 promoter. The fragments encompassing nucleotides 17–597, 577–1182 and 1162–1696 of the β TrCP1 cDNA coding region were amplified by PCR with a *Pfu* Turbo DNA polymerase (Stratagene) and subcloned into pcDNA3.1 downstream of the T7 promoter.

The expression vectors for β -catenin^{S33Y}, Tcf4 (wild type) and dnTcf4 were gifts from B. Vogelstein and K. Kinzler. The Tcf1 expression vector was obtained from ATCC. IMP2 and IMP3 were provided by D. Herrick. pSV2 was from ATCC. The pSilencer 1.0-U6 short interfering RNA (siRNA) expression vector was from Ambion. Flag-CRD-BP was constructed as described previously¹¹. Construction of CRD-BP shRNA was performed in accordance with the recommendations of Ambion. In brief, siRNA sequences were selected with the siRNA Target Finder and Design Tool (http://www.ambion.com/techlib/misc/siRNA_finder.html). The annealed shRNA insert was cloned into the pSilencer 1.0-U6 siRNA expression vector.

Antibodies and immunotechniques. Mouse anti-Flag M2 antibody (Sigma-Aldrich), antibodies against β -actin (Santa Cruz), β -catenin (Cell Signalling) and I κ B α (Cell Signalling) were purchased. Anti-CRD-BP antibody was generated as described previously²⁰. Secondary antibodies conjugated with horseradish peroxidase (Chemicon) were purchased. Immunoprecipitation and immunoblotting procedures were as described elsewhere⁵.

Tissue culture and transfections. 293T cells maintained in DMEM medium with 10% FBS were transfected with the calcium phosphate method. Colon cancer cell lines HCT116 and Caco2 were cultured in RPMI medium with 10% FBS and transfected with Lipofectamine 2000 (Invitrogen), in accordance with the manufacturer's recommendations. The amount of DNA in each transfection was kept constant by the addition of an appropriate amount of empty expression vector.

β TrCP1 mRNA degradation in vivo. 293T cells were transfected with 5 μ g of Tet-off plasmid (Clontech), 3 μ g of pBIG- β TrCP1 and β -catenin^{S33Y}/Tcf4 expression vectors (3 μ g each) or 4 μ g of FLAG-CRD-BP per 100-mm plate. At 48 h after transfection, the cells were treated with doxycycline (1 μ g ml⁻¹) and harvested at different time points after treatment. The levels of β TrCP1 mRNA were analysed by northern blot analysis with a β TrCP1-specific probe.

Whole cell extracts and other assays. To analyse the turnover of β -catenin and I κ B α , the cells were lysed in non-denaturing lysis buffer (10 mM HEPES pH 7.6, 3 mM MgCl₂, 40 mM KCl, 2 mM dithiothreitol, 5% (v/v) glycerol, 0.5% (v/v) IGEPAL, and protease inhibitors). To obtain the whole cell lysate for all the other experiments, the cells were lysed with RIPA buffer. Ribosomal salt wash (RSW) was prepared as described previously²⁸. Northern blot analysis and run-on assay were performed as described previously⁵. Analysis of RNA–protein complexes is described in Supplementary Methods.

Primary human tissue samples. Protein extracts were prepared from tissue samples obtained in the course of direct surgery from nine patients with primary colorectal carcinomas in the Cancer Research Institute, Kanazawa University, Kanazawa, Japan (described in ref. 18). From each patient, tissue samples were taken from non-neoplastic mucosa at the proximal surgical margin and the primary tumour (at T3 stage).

Received 3 November 2005; accepted 20 April 2006.

- Polakis, P. The oncogenic activation of β -catenin. *Curr. Opin. Genet. Dev.* **9**, 15–21 (1999).
- Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11–23 (2001).

- Wilusz, C. J., Wormington, M. & Peltz, S. W. The cap-to-tail guide to mRNA turnover. *Nature Rev. Mol. Cell Biol.* **2**, 237–246 (2001).
- Eastman, Q. & Grosschedl, R. Regulation of LEF-1/TCF transcription factors by Wnt and other signals. *Curr. Opin. Cell Biol.* **11**, 233–240 (1999).
- Spiegelman, V. S. et al. Wnt/ β -catenin signaling induces the expression and activity of β TrCP ubiquitin ligase receptor. *Mol. Cell* **5**, 877–882 (2000).
- Nielsen, J. et al. A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol. Cell. Biol.* **19**, 1262–1270 (1999).
- Nielsen, F. C., Nielsen, J. & Christiansen, J. A family of IGF-II mRNA binding proteins (IMP) involved in RNA trafficking. *Scand. J. Clin. Lab. Invest. Suppl.* **234**, 93–99 (2001).
- Runge, S. et al. H19 RNA binds four molecules of insulin-like growth factor II mRNA-binding protein. *J. Biol. Chem.* **275**, 29562–29569 (2000).
- Atlas, R., Behar, L., Elliott, E. & Ginzburg, I. The insulin-like growth factor mRNA binding-protein IMP-1 and the Ras-regulatory protein G3BP associate with tau mRNA and HuD protein in differentiated P19 neuronal cells. *J. Neurochem.* **89**, 613–626 (2004).
- Bernstein, P. L., Herrick, D. J., Prokipcak, R. D. & Ross, J. Control of c-myc mRNA half-life in vitro by a protein capable of binding to a coding region stability determinant. *Genes Dev.* **6**, 642–654 (1992).
- Doyle, G. A. et al. The c-myc coding region determinant-binding protein: a member of a family of KH domain RNA-binding proteins. *Nucleic Acids Res.* **26**, 5036–5044 (1998).
- Prokipcak, R. D., Herrick, D. J. & Ross, J. Purification and properties of a protein that binds to the C-terminal coding region of human c-myc mRNA. *J. Biol. Chem.* **269**, 9261–9269 (1994).
- Tessier, C. R., Doyle, G. A., Clark, B. A., Pitot, H. C. & Ross, J. Mammary tumor induction in transgenic mice expressing an RNA-binding protein. *Cancer Res.* **64**, 209–214 (2004).
- Kudo, Y. et al. Role of F-box protein β Trcp1 in mammary gland development and tumorigenesis. *Mol. Cell. Biol.* **24**, 8184–8194 (2004).
- Fuchs, S. Y., Spiegelman, V. S. & Kumar, K. G. The many faces of β -TrCP E3 ubiquitin ligases: reflections in the magic mirror of cancer. *Oncogene* **23**, 2028–2036 (2004).
- Roose, J. et al. Synergy between tumor suppressor APC and the β -catenin-Tcf4 target Tcf1. *Science* **285**, 1923–1926 (1999).
- Morin, P. J. β -Catenin signaling and cancer. *BioEssays* **21**, 1021–1030 (1999).
- Ougolkov, A. et al. Associations among β -TrCP, an E3 ubiquitin ligase receptor, β -catenin, and NF- κ B in colorectal cancer. *J. Natl Cancer Inst.* **96**, 1161–1170 (2004).
- Lin, A. & Karin, M. NF- κ B in cancer: a marked target. *Semin. Cancer Biol.* **13**, 107–114 (2003).
- Ross, J., Lemm, I. & Berberet, B. Overexpression of an mRNA-binding protein in human colorectal cancer. *Oncogene* **20**, 6544–6550 (2001).
- He, T. C. et al. Identification of c-MYC as a target of the APC pathway. *Science* **281**, 1509–1512 (1998).
- Hansen, T. V. et al. Dwarfism and impaired gut development in insulin-like growth factor II mRNA-binding protein 1-deficient mice. *Mol. Cell. Biol.* **24**, 4448–4464 (2004).
- Leeds, P. et al. Developmental regulation of CRD-BP, an RNA-binding protein that stabilizes c-myc mRNA in vitro. *Oncogene* **14**, 1279–1286 (1997).
- Ioannidis, P. et al. CRD-BP: a c-Myc mRNA stabilizing protein with an oncofetal pattern of expression. *Anticancer Res.* **23**, 2179–2183 (2003).
- Ioannidis, P. et al. Expression of the RNA-binding protein CRD-BP in brain and non-small cell lung tumors. *Cancer Lett.* **209**, 245–250 (2004).
- Ioannidis, P. et al. C-MYC and IGF-II mRNA-binding protein (CRD-BP/IMP-1) in benign and malignant mesenchymal tumors. *Int. J. Cancer* **94**, 480–484 (2001).
- Ioannidis, P. et al. 8q24 Copy number gains and expression of the c-myc mRNA stabilizing protein CRD-BP in primary breast carcinomas. *Int. J. Cancer* **104**, 54–59 (2003).
- Ross, J. Messenger RNA turnover in cell-free extracts from higher eukaryotes. *Methods Mol. Biol.* **118**, 459–476 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Spiegelman for help with the manuscript preparation. This work was supported by an American Cancer Society Award (to V.S.S.). The work was supported in part by a University of Pennsylvania Cancer Center Pilot Grant and an NCI grant (to S.Y.F.), by NIH grants (to J.R.) and by the Japanese Ministry of Education, Science, Sports, Technology and Culture, by the Ministry of Health, Labor and Welfare, and by the Japan Society for the Promotion of Science (to T.M.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.S.S. (spiegelman@dermatology.wisc.edu).

CORRIGENDUM

doi:10.1038/nature04909

A brain-specific microRNA regulates dendritic spine development

Gerhard M. Schratt, Fabian Tuebing, Elizabeth A. Nigh, Christina G. Kane, Mary E. Sabatini, Michael Kiebler & Michael E. Greenberg

Nature 439, 283–289 (2006)

The protocol for penetratin-coupling of siRNA-like molecules used in this Article incorporated suggestions by Carol M. Troy that were based on the method described in ref. 1.

1. Davidson, T. J. *et al.* Highly efficient small interfering RNA delivery to primary mammalian neurons induces microRNA-like effects before mRNA degradation. *J. Neurosci.* 24, 10040–10046 (2004).

CORRIGENDUM

doi:10.1038/nature04910

Brain-state- and cell-type-specific firing of hippocampal interneurons *in vivo*

Thomas Klausberger, Peter J. Magill, László F. Márton, J. David B. Roberts, Philip M. Cobden, György Buzsáki & Peter Somogyi

Nature 421, 844–848 (2003)

In the Methods section of this paper, the fifty-two male Sprague–Dawley rats were incorrectly described as having been anaesthetized with 1.25 mg kg^{−1} urethane. The actual dose of urethane was 1.25 g kg^{−1}.

CORRIGENDUM

doi:10.1038/nature04911

Universal scaling of respiratory metabolism, size and nitrogen in plants

Peter B. Reich, Mark G. Tjoelker, Jose-Luis Machado & Jacek Oleksyn

Nature 439, 457–461 (2006)

It has been drawn to our attention that the wording used in the first sentence after equation (1) is ambiguous. For the example used to illustrate $\frac{3}{4}$ power scaling, we implied that if metabolic rate had a body mass scaling exponent of 0.75, a 10-fold increase in size would increase metabolic rate 7.5-fold; however, we were referring to the power-law relationship on a log scale, which should have been clear from the context. A 10-fold increase in mass would therefore increase metabolic rate as a function of $10^{0.75}$, which is equal to a 5.62-fold increase in metabolic rate. The apparent error in this example has no bearing on the power-law relationship or its interpretation in the paper.

naturejobs

**THE CAREERS
MAGAZINE FOR
SCIENTISTS**

Getting away from science is next to impossible. Regardless of whether you labour in the lab or, like me, you slave over a keyboard writing about research, the next story or experiment sticks with you even when you're away from your place of work. But sometimes, a little distance, a change of scenery — and freedom from phone calls and e-mail — can help to sharpen your focus.

A group of friends and I achieved both distance and clarity during a recent trek through the Grand Canyon. Hiking 80 kilometres through the desert, while carrying 15-kilogram packs and seeing more animals than people, has a way of paring life down to its essentials. Even so, getting away from science still proved to be extraordinarily difficult — especially as everyone in our party had some kind of connection to research. My wife Polina is a social scientist and policy analyst, Matt trained as a forester, Kate originally qualified as an entomologist, and Renee and Eric worked for the National Oceanic and Atmospheric Administration.

To our credit, we didn't talk about work — much. We were too busy locating the next water source, dodging rattlesnakes and gazing at bighorn sheep. But our respective professional perspectives still came into play. Matt and Eric's experience with maps, compasses and the global positioning system kept us heading in the right direction. Renee and Kate's skills at plant and insect identification made us more aware of the flora and fauna we encountered. And Polina's knack for getting on with people scored us some extra food from other hikers who wanted to leave their excess weight behind for their ascent.

So I couldn't help thinking how our group's skills and professional paths reflect a microcosm of scientific careers. Only one of the six is even considering a traditional career in research. But we all still draw on our scientific backgrounds in our work and daily life. And using our soft skills in communication, planning and management made our trip both safe and memorable — and left us more focused on our jobs when we returned.

Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Gene Russo

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
Peter Bless
Classified Sales Representatives
Tel: +1 800 989 7718
East USA/Canada: Andrew Bennie
**NIH/Maryland/New York/
Pennsylvania:** Shelley Cohen

San Francisco Office
Classified Sales Representative:

Michaela Bjorkman
West USA/West Corp. Canada
225 Bush Street, Suite 1453
San Francisco, CA 94104
Tel: +1 415 781 3803
Fax: +1 415 781 3805
e-mail: m.bjorkman@naturesf.com

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
Advertising Production Manager:
Stephen Russell
To send materials use London
address above.
Tel: +44 (0) 20 7843 4816

Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
Naturejobs web development:
Tom Hancock
Naturejobs online production:
Catherine Alexander

European Satellite Office
Patrick Phelan
e-mail: p.phelan@nature.com

Japan Head Office, Tokyo
Chiyoda Building,
2-37 Ichigayatamachi,
Shinjuku-ku,
Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
e-mail: a.watanabe@natureasia.com

MOVERS

Nicholas Schork, director, Center for Biomedical Informatics; co-director, Cancer Genetics Program, University of California, San Diego



2001-present: Professor of psychiatry and biostatistics, University of California, San Diego

1997-2001: Associate professor of epidemiology and biostatistics, Case Western Reserve University, Cleveland, Ohio

1999-2000: Vice-president of statistical genomics, Genset, La Jolla, California

An philosopher at heart, Nicholas Schork was an unlikely future director of a cancer programme. To make ends meet, the philosophy graduate student tapped into his more lucrative skills as a computer programmer for the University of Michigan's medical school. In doing so, he began a circuitous route towards becoming a geneticist.

Philosophers have a tradition of being inspired by great teachers. Schork's interest in genetics and genomics was fuelled by luminaries such as Francis Collins, now director of the National Human Genome Research Institute but back then a professor at the University of Michigan. As Schork's interest in biology grew, so did his desire to address the theoretical problems in genomics. His father, a biostatistician, encouraged him to find mentors in biostatistics and epidemiology.

Along with his MA in philosophy, Schork began a PhD in statistical genetics. He had no idea how much the discipline would permeate science. "I thought I'd write my thesis in obscurity and resign myself to a basement office in a far corner of a university teaching biostatistics," he says.

As it turned out, he couldn't have been in a better position to capitalize on what, at the time, was a rare set of skills. Bypassing a postdoc stint, he took an assistant professorship at Case Western Reserve University (CWRU) in Cleveland, Ohio. A colleague introduced him to biologist Eric Lander, founding director of the Broad Institute in Cambridge, Massachusetts — one of several serendipitous meetings that accelerated Schork's career. Together, he and Lander wrote a landmark review of the genetic underpinnings of complex traits.

Linking genetic variation to disease risk raised so many questions that Schork took leave from CWRU to explore the potential for a map of linked genetic markers, or haplotype, for the French company Genset, one of the first players in the field. The project was completed by the public-private International HapMap Consortium, and Schork then focused on integrated statistical approaches to understanding the genetic basis of disease. When the University of California, San Diego, offered a chance to help establish a genetics/genomics presence, he jumped at it.

In his dual roles as director of the Center for Biomedical Informatics and co-director of the Cancer Genetics Program, Schork proves that abstract thinkers have a significant role to play in science. "I'm still a student all these years later," he says.

Virginia Gewin

RECRUITERS & ACADEMIA

A physics walkabout

I'm living a great postdoctoral adventure. On a 27-month fellowship funded by the US National Science Foundation (NSF), I will spend time at several institutions, with the University of Cambridge, UK, as my base. It's an amazing opportunity. I get plenty of money for living and travelling, a fancy title, ample freedom and the chance to work with famous professors on three continents — bolstering not only my résumé but also my personal growth.

I'm on a Maths and Physical Sciences Distinguished International Postdoctoral Research Fellowship, which is aimed at enhancing connections between the US science and engineering community and its international counterparts.

As a graduate student in condensed-matter physics at the University of Wisconsin, Madison, I worked as a theorist on the use of electrons in silicon nanostructures for quantum information processing.

The itinerant nature of my fellowship could make productivity a challenge. Fortunately, condensed-matter and quantum-information physics — with their tendency towards self-contained papers with few authors — are well suited to far-flung collaborations. In fact, work in emerging areas often benefits from interdisciplinary, cross-institutional cooperation.

My first stop was the University of Melbourne in Australia, from which I have just returned after four months. It was a natural choice: we in Madison have good relations with the Australians, whose government has invested heavily in a Centre for Quantum Computer Technology. Plus, I have family there.

Australia's small physics community feels somewhat isolated. Save for the work of a well-known few, the fact that good physics happens there is often overlooked. The isolation and the relatively short PhD programme (3.5 years) can make it hard to compete for postdocs abroad.

Melbourne was wonderful. I often hung out with my cousins on sunny afternoons. I tackled projects in silicon, quantum optics and diamond devices, whose properties make them promising for quantum-technology applications. Especially exciting was the chance to work with experts in other subfields such as atomic physics. I have fond memories of group meetings over beers at the university pub.

Now I'm 'home' at the legendary Cavendish Laboratory. Soon I'll be off for a couple of months in Japan. I'll be writing about those experiences soon.

Charles Tahan is an NSF distinguished research fellow.

GRADUATE JOURNAL

Graduation joy

The completion of my PhD marks the official end of my student life, which will constitute close to 75% of my lifetime by the time I graduate. I'll finally move on after years of homework and exams, not to mention many hours of research on the effects of rising atmospheric carbon dioxide on carbonate ecosystems.

Completion gives satisfaction. To finally conquer a seemingly endless challenge, which right now seems to be the project of my life, is a source of tremendous joy. At first, I didn't see the benefit of receiving a fancy piece of paper with three rather ordinary letters after my name. But then I remembered working around the clock — in the field collecting water samples, in the lab analysing chemical properties, and in front of the computer creating numerical models, typing papers and analysing my results. And, of course, there was the amount of free time that I spent contemplating and reflecting on my research. Suddenly that piece of paper seems invaluable and those three letters magical.

Eventually, the anticipated graduation euphoria will subside and I may come to realize that it was just another stepping stone. Around the corner, adventures are lining up. With a few months to go, I still do not know what route I'll choose. It may be a postdoc, a teaching position, something outside academia, or I may even pack my bags and explore the world for a while. Regardless, I'm excited by the prospect of myriad possible directions.

Andreas Andersson is a final-year PhD student in oceanography at the University of Hawaii.

Great unreported discoveries no. 163

It's good to talk...

Mike Resnick

When the team at Iowa State (or was it Nebraska?) came up with proof positive that plants feel pain, it made headlines not just that day, but for months thereafter. We have millions of people who became vegetarians because they didn't want animals to die just so they could fill their stomachs, and suddenly they discovered that everything they eat feels pain.

I found it fascinating. That's why I changed my major and went into botany — because I couldn't stop wondering: if plants can feel, what else can they do? Like, for example, can they think?

Of course, thinking in itself is a dead end, especially if you're rooted to one spot, unless you can communicate your thoughts, so that's what I really tried to specialize in. It's a good thing I knew how to get government grants, because I spent the first 14 years after getting my PhD without any hint of success.

I would talk to them. I would play them music. I would write messages in every known language and hold them up. I even brought in professors who could speak dead languages. All to no avail.

Undaunted — well, not very daunted anyway — I brought in psychics to see if they could form a bond with any of the plants in my lab. Still no luck.

I would put bees and butterflies in screened cages and explain to the flowers that if they wanted to reproduce, all they had to do was tell me, and I would release the insects and the process of regeneration could get under way. Nothing.

Finally I tied them into computers, exceptionally bright machines that could turn almost any signal, no matter how slight, how basic, how weak, how alien — into a spoken translation. All I got was silence.

I still remember the breakthrough. I'd just met and lost my heart to Bubbles La Tour, a truly wonderful dancer who, despite her billing, could hardly be called a stripper as she started out naked (and then got energetic). I walked over to one of the hybrid daisies in the lab and began plucking off its petals one by one, muttering "She loves me...she loves me not...she loves me..."

"Ouch!" said a strange voice.

I looked around the lab, but I couldn't see anyone.

After a moment I decided I had imag-

ined it, and I pulled off another petal.

"Damn, that smarts!" said the voice again. "What did I ever do to you?"

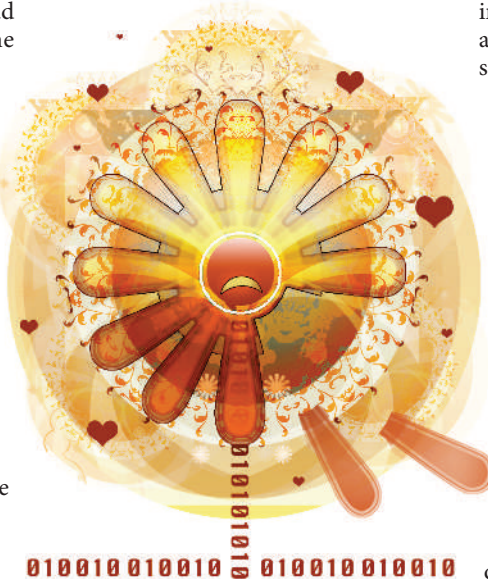
"I beg your pardon?" I said, looking around and trying to spot the speaker.

"Begging my pardon is all very well and good," said the voice. "But you're denuding me. Are you going to disembowel me next?"

"Who said that?" I demanded.

"Whose limbs have you been pulling off?" the voice shot back.

I stared at the daisy, and suddenly I saw that it was still attached to the computer. I'd run my most recent experiment two days



before and hadn't remembered to disconnect it.

"You?" I said, bending over it.

"Yes, me," said the daisy. "And brush your teeth if you're going to stand this close to me. I'm supposed to be living in a world of gorgeous scents."

"You can talk!" I exclaimed excitedly, and then repeated: "You can talk!"

"What a stunning observation," said the daisy. "You must have been the brightest one in your class."

"There's no need for sarcasm," I said.

"There's no need for sadism, either, but you kept pulling off my limbs."

"Your petals," I corrected it.

"Semantics," said the daisy.

"Well, now that I know you can talk, I promise never to do it again," I said.

"I don't want to seem ungracious, but how long have you known I could feel pain?"

"It got you to speak, didn't it?" I said defensively.

"So when your spouse is mad at you and won't talk, do you pull off her arms and legs?"

"Hardly ever," I admitted.

"Well, there you have it."

"Look, I apologize, all right?" I said. "This is the most historic breakthrough in the history of the human race! We should be celebrating!"

"Well, it's not the most historic breakthrough in the history of the daisy race," the flower replied. "I find you arrogant, self-centred, and a rather dull conversationalist. I am all through communicating with human beings now and forever, and I'm going to go back to peacefully and silently contemplating my navel."

"You don't have a navel," I pointed out.

"I was couching the concept in terms you'd understand, which was clearly a waste of time. Now please go away."

I spoke to it for another half an hour, but couldn't get any response. I considered calling in Doctors DiChario and Gormley, but then I did some serious thinking, and I decided that if the daisy kept its promise and didn't speak to them, they'd think I was either crazy or a liar (or both). And if it did speak to them, they'd apply for the same grants I was living on, and with their superior credentials they'd wrest them away from me and I'd have to go out and find another (and doubtless more difficult) way to make a living. If I tried to make my findings public and the daisy passed the word to its kith and kin, no one would be able to verify it and I'd become a laughing stock. And if the other flowers were as foul-tempered and obnoxious as the daisy, why would anyone want to communicate with them in the first place?

The more I thought about it, the more I decided that nothing is sometimes the very best thing to do.

I carefully disconnected the daisy from the computer.

Some time later I found myself thinking about Bubbles La Tour again, and my hand absently went to the flower.

"She loves me..." I intoned dreamily, "She loves me not..."

Five-time Hugo winner Mike Resnick is the author of more than 50 novels and 200 stories, and has edited more than 40 anthologies. He has won major awards in six countries, and his work has been translated into 22 languages.